VIP: A UNIFYING FRAMEWORK FOR COMPUTATIONAL EYE-GAZE RESEARCH

MA KENG TECK

M.Sc (CompSci), NUS, 2002

B. Sci. (Hons), NUS, 2001

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Ma Keng Teck
October 9, 2014

Acknowledgment

Firstly, I would like to offer my sincere and deepest gratitude to my advisor Professor Terence Sim. His immense knowledge, patience and insights had helped me greatly in my Ph.D. study. I always came away from our discussions more enlightened.

Besides my advisor, I would like to thank the rest of my thesis committee: Professor Low Kok Lim and Professor Yan Shuicheng, for their critical feedback and insightful questions. My thesis is so much better because of them.

I would also like to thank Professor Mohan Kankanhalli. By offering me the Research Assistant position, he had truly opened a new door for me when another closed. I can always depend on his sage advices for so many things, both for my research and my personal life. My Ph.D. journey will be a lot tougher without his unwavering support and guidance.

A special thanks goes to A*STAR, Agency for Science, Technology and Research for sponsoring part of my research.

I would like to thank my fellow lab mates, especially Hamed, Hossein, Karthik, Lewis, Vlad and Zhang Li. It was hard fun we had, research is never done.

Last but not the least, I would like to thank my family. Especially, the VVIP in my life, my wife.

October 9, 2014

Contents

Name	:	Ma Keng Teck
Degree	:	Doctor of Philosophy
Supervisor(s)	:	Associate Professor Sim Mong Cheng, Terence
Department	:	Department of Computer Science
Thesis Title	:	VIP: A unifying framework for computational eye-gaze research

Summary

Eye-gaze data has been used in wide range of computer science research. For saliency research, the reference model is bi-directional: top-down and bottom-up. In biometric research, the identity of a person can be inferred from eye-gaze. In human-computer interface, eye-gaze is a response of the interactions between the tasks and the visual stimulus. These models are incomplete and we propose the VIP framework. This formal framework captures the dependence of eye-gaze on Visual stimuli, Intent, and Person, making it more complete and subsuming all existing models.

We had conducted extensive user experiments to collect the VIP and VVIP datasets. These are the first eye-gaze datasets to contain all 3 VIP factors, for images and videos respectively. For the benefits of the research community, these datasets are made publicly available at http://mmas.comp.nus.edu.sg/VIP.html.

The utility of the VIP framework is illustrated with 2 novel problems: (i) inferring the viewer's personal traits and (ii) an implicit just-in-time profiling system: `Eye-2-I`. The first application is a novel use of eye-gaze data; the second is the first system to fully encapsulate the 3 factors of the eye-gaze data for comprehensive profiling of a user.

We have also developed a novel feature extraction algorithm, `ROI`, which is modeled after the human's foveal vision. The features extracted are useful for inferring the interests of the subjects as shown by our experimental results.

Finally, we demonstrate the superiority of our framework over existing models by proposing a trait-specific fixation prediction approach, which has higher AUC scores than current trait-agnostic approaches for some images.

Keywords : Eye-tracking, Computational model, User Profiling, Personalisation, Dataset, Saliency

List of Tables

List of Figures

Chapter 1

Introduction

```
The real voyage of discovery consists
not in seeking new landscapes
but in having new eyes.

- Marcel Proust (Author)
```

"The eyes are the windows to the soul," goes an old English proverb. We agree. In fact, the window acts both ways: as a portal into the person's mind, and as a lens to perceive visual stimuli. In this regard, eye-gaze — the coordinated motion of the eyes and the head — can provide invaluable clues both to the viewer, and to the object being viewed. This is the exciting premise, and promise, of research using eye-gaze data. Eye-gaze not only permits a fresh approach to existing problems (such as image segmentation), but also throws open a brave new world in which new applications may be created, and new inferences made.

The importance of eye-gaze data has been growing in the past few years due to a confluence of several factors: better equipment, greater adoption of eye-gaze by the computer science research community, and the increasing awareness that eye-gaze "imprints" many hidden psychological and perceptual aspects of the viewer's mind.

1.1 Better equipment

In the past decade, eye-gaze was not used much because of the high cost of sensors and their unwieldy size. But recent advances in eye-tracking technology have produced cheaper, more reliable and smaller trackers. Thus, more eye-gaze data has become available for research, and many new applications are now possible. One good example is the Samsung Galaxy S 4 cell phone (Samsung, 2013). It boasts an eye-tracking sensor embedded in a compact, battery-operated system at a competitive price of approximately US$600. Likewise, Tobii is also releasing an affordable (US$995) desktop eye-tracker by the end of 2013 (Tobii, 2013). Besides cost, ease-of-use and tracking precision have also improved. This is encouraging researchers, previously put off by complicated devices, to incorporate eye-gaze as an additional modality into their experiments. In turn, these researchers are publicly releasing more eye-gaze datasets (Winkler and Subramanian, 2013), which can only advance the field (Borji and Itti, 2013) further.

1.2 Greater adoption

In recent years, researchers in multimedia and computer vision are turning to eye-gaze data to better solve traditional problems in their fields (Ramanathan et al., 2010; Katti et al., 2011; Ramanathan et al., 2009; Jaimes et al., 2001; Yadati et al., 2013; Vural and Akgul, 2009; Bulling et al., 2011). Newer research topics have also arisen, such as visual saliency in stereoscopic images (Lang et al., 2012). Frintrop et al.'s survey is an excellent read for a more-depth discussion of computational visual attention systems from a cross-disciplinary point of view (Frintrop et al., 2010).

1.3 Psychological imprint

There is increasing evidence that eye-gaze bears the imprint of several types of psychological and perceptual factors (Martinez-Conde et al., 2004; Hoffman and Subramaniam, 1995). For example, Chua et al. (2005)'s experiments showed that people from different cultural background have differing eye-gaze patterns. This makes eye-gaze data very useful for inferring high-level semantics in multimedia processing. Compared to other modalities, such as textual tags, eye-gaze provides a more direct route to infer the viewer's mental state because it is tightly coupled with the visual stimuli. This has, in fact, been exploited to aid the localization and labeling of images (Ramanathan et al., 2009). Eye-gaze also contains information of the stimuli such as saliency, emotive content, objects and their relationships (Ramanathan et al., 2010). Furthermore, it can be used to infer a person's emotions (Katti et al., 2011), fatigue (Schleicher et al., 2008), identity (Rigas et al., 2012) and tasks (Bulling et al., 2011).

1.4 Unifying framework

We further advance eye-gaze research by inventing a unifying formal framework with which to reason about eye-gaze. We review existing computational models used in current eye-gaze research, and show that, while they are appropriate for their given applications, they are, alas, incomplete. We then propose our VIP eye-gaze framework, which captures the dependence of eye-gaze on Visual stimuli, Intent, and Person. By *visual stimuli* we include any visual modality, such as traditional images and videos, and also novel mediums like 3D images and games. By *intent* we refer to the immediate state of the mind such as purpose of viewing the stimuli, the emotions elicited by the stimuli, etc. For example, one can view an image to count the number of people in it, or one can simply be appreciating

the image as an art form. This difference in purpose produces different eye-gaze patterns. Finally, by *person* we mean the persistent traits of the viewer of the visual stimuli, including identity, gender, age, and personality types (See Chapter 3.1 for formal definitions). Because we have done careful survey of the field, we believe our formal VIP framework is more complete, and subsumes all existing eye-gaze models.

We illustrate the utility of our framework with two applications: inferring the demographic and personality traits of the viewer; and just-in-time and implicit user profiling from eye-gaze. We demonstrate the superiority of our framework over existing models by proposing a trait-specific fixation prediction approach.

This substantiates our claim in the first paragraph of this chapter: that the eye-window works in both directions. As far as we can tell, this inference of personal traits for eye-gaze is pioneering; no one else has done this before. Indeed, the reader will quickly see, that given our VIP framework, many new research opportunities lie just ahead. Our hope is by providing a more complete and unified formal framework for computational eye-gaze research, the advances in this exciting field will accelerate even faster.

With the VIP framework as a map, I would like to invite the reader to join me in voyage of discoveries for computational eye-gaze research.

1.5 Contributions

Some research improves on solutions for existing problems and others propose interesting novel problems. And a rare few impact their respective fields as a whole by making new fundamental insights or discoveries. In this thesis, we make contributions to all these categories.

Our contributions are as follows:

1. **Unifiying framework**: Our VIP framework unifies the current com-

putational models in eye-gaze research. It is the first to identify and highlight the importance of the P factors. It standardizes comparisons and formulations of existing and new computational eye-gaze research and applications (Chapter 3). Furthermore, it allows differences to be systematically examined; assumptions to be formally reviewed and gaps in research direction to be more thoroughly investigated (Chapter 8).

2. **Eye-gaze datasets with Personal data**: Our VIP dataset is the first eye-gaze dataset to include all 3 VIP factors. Our Video-VIP (VVIP) dataset is the first multi-modal dataset (facial expression, eye-gaze, video affects analysis and text) coupled with anonymous demographic profiles, personality traits and topics of interest (Chapter 4).

   By applying existing algorithms to these datasets, the algorithms' assumptions can be validated or rejected with the reference of the complete VIP factors. New algorithms and novel research problems are also made possible with the availability of these datasets. We show the utility of the VIP dataset by proposing a novel research problem (Chapter 5) and improving on an existing problem (Chapter 6) with it. The VVIP dataset was used to validate our implicit just-in-time profiling system.

3. ***Implicit* and *just-in-time* demographic and personality profiling**: We are the first to use eye-gaze when viewing images to implicitly profile people's demographic and personality traits just-in-time. The eye-gaze modality offers several advantages over other modalities (Chapter 5).

   Our proposed `Eye-2-I` system is able to infer detailed user profile: demographic, personality types, interests and emotions from the user's eye-gaze viewing videos implicitly and just-in-time (Chapter 7).

4. **Improvements in fixation prediction**: Our proposed P-specific fixation prediction approach is more accurate than the current P-agnostic approach for some images (Chapter 6).
5. **Novel feature extraction algorithm for inference of personal traits and interests from eye-gaze**: Our novel features, `ROI` outperforms the statistical features for inferring of interests in video-watching activity (Chapter 7).

The framework and VIP dataset were published in the peer-reviewed 4th International Workshop on Human Behavior Understanding (Ma et al., 2013). The `Eye-2-I` system is currently under peer-review for the SIG CHI conference and has been filed for Invention Disclosure to the university. See Appendix B for the complete listing of my publications and pending patent application.

Chapter 2

Background

`Study the past if you would define the future.`

`- Confucius`

The first eye-tracker was invented in 1878. While it was intrusive, bulky and expensive, many scientific discoveries were enabled by it and its successors. From then on, eye-tracking technology has progressed by leaps and bounds. By 1981, the first real-time interactive eye-tracking system was developed. This breakthrough had made many new applications possible. In 2013, a battery operated eye-tracker is embedded into a 130 grams smart-phone which costs US$600. More technological breakthroughs in eye-tracking are expected as the modality began to reach the mass consumer market.

The range of devices with eye-tracking capabilities is ever-increasing. Broadly, there are 3 types of eye-trackers: fixed, mobile and wearable. Fixed eye-trackers are commonly found integrated with larger displays, such as smart TV and desktop computers (Tobii, 2013). Mobile eye-trackers are integrated with devices such as smart-phones and laptops (Samsung, 2013; Amazon, 2014). Wearable eye-trackers are head-mounted and used for recording of eye-gaze in natural and unconstrained settings (SMI, 2014).

Eye-trackers can also be used either indoors or outdoors.

In the past, reliable eye-trackers cost more than US$10,000, severely limit access of this modality to the wider research community. With the availability of affordable but high quality eye-trackers, such Tobii PC-Eye Tobii (2013), (US$995), it becomes increasingly clear that eye-gaze research will be accessible for a much larger group of researchers. In fact, the number of public eye-gaze datasets has been growing at a steady rate (Chapter 4).

To summarize, here's a brief history of eye-trackers (Jacob and Karn, 2003):

- 1878: First eye-tracker. Intrusive.
- 1901: First non-intrusive eye-tracker. Horizontal only.
- 1905: First 2-dimensional eye-tracker.
- 1948: First head-mounted eye tracker.
- 1973: First non-obtrusive eye-tracker.
- 1981: Real-time interactive eye-tracking system.
- 2011: Tobii PCEye. US$6900. 250g.
- 2013: Samsung Galaxy S IV. 130g. Battery operated. Embedded. US$600.
- 2013: Tobii REX. End 2013. US$995.
- 2014: SMI Eye-tracking Glasses 2, wireless. US$11,900.
- 2014: Amazon Fire Phone. US$449.

Most current generation eye-trackers record the locations and timestamps of the human subject's gaze at fixed time intervals. Other data such as the pupillary dilations and blinks may also be recorded. These data are then pre-processed into fixations (locations and durations) and saccades (sequences and velocities). The ordered sequences of fixations and saccades are referred to as scanpath. The fixation and saccade patterns provide valuable insights to the visual attention of the subjects (Martinez-

Conde et al., 2004; Hoffman and Subramaniam, 1995). The saccades, pupillary dilations and blinks reveal their mental states such as emotions and fatigue (Schleicher et al., 2008; Bradley et al., 2008). The scanpath data is mostly used by usability studies and sometimes for biometric research. Figure 2.1 illustrates the fixations, saccades and scanpaths data.

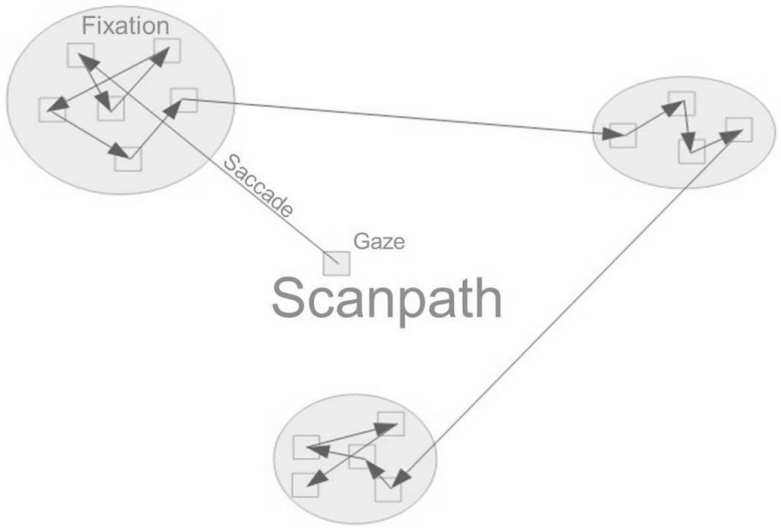

Figure 2.1: Illustration of fixations, saccades and scanpath. Each arrow represents a saccade: a ballistic movement of gaze. The squares are the pauses between the saccades. During these pauses, information about the stimulus is processed by the brain. The circles denote the fixations: clusters of consecutive pauses. Saccades which are within a fixation is known as micro-saccades. The inter-fixations saccades are called exploratory or orienting saccades. Scanpaths are typically defined as the ordered sequence of fixations and orienting saccades; and infrequently as the ordered sequence of fixations and all saccades.

2.1 Eye-tracker calibration

The output of eye trackers, typically, position of the pupil etc., needs to be mapped to the coordinates of the stimuli, e.g. screen. This process is known as *calibration* procedure. During this procedure, the eye tracker measures characteristics of the user's eyes and maps them to the known coordinates of pre-defined area/points of the stimuli. This mapping may include information about shapes, light refraction and reflection properties

of the different parts of the eyes (e.g. cornea, placement of the fovea, etc.). During the calibration the user is asked to look at pre-defined points on the stimuli, also known as calibration dots. During this period measurements of the eyes are collected and analyzed. The resulting information is then mapped to calibration points. There may be a further validation step to determine the errors of the calibration.

2.2 Current eye-trackers

Current technology can track the eye-gaze at up to $1250Hz$, with accuracy of $0.25°$ to $0.5°$. Latency is less than 0.5 milliseconds (SensoMotoric Instruments, 2013). These impressive technological achievements allow for both real-time and high precision applications. On the other end of the spectrum are affordable open-source softwares using the webcam's as the input source (CodeProject, 2013). This allows for low cost, web-based and crowd-sourced research studies. The quest for more accurate and robust eye-tracking systems is still an ongoing endeavor (Hansen and Ji, 2010). It will also soon be possible to have general purpose devices which can also track eye-gaze such as smart phones or wearable devices (Arnon Amir et al., 2003; Hodge and Rosenblatt, 2013). Another eye-tracking device is the electrooculography (EOG) system. While the EOG is somewhat obtrusive as it needs to be in contact with the subject, it is very lightweight and works under different lighting conditions and can be worn as an embedded, self-contained system. It is most suitable for measuring eye movement in mobile daily-life situations. As such, researchers have the flexibility to choose the most suitable device accordingly, making eye-tracking an attractive modality.

In many ways, eye-trackers are like Global Positioning System (GPS) 10 years ago. Before the immense popularity of smartphones, GPS tech-

nology was maturing and used in many specialized domains. However, there was very little research in geolocation applications. Today, with the prevalence of GPS-equipped smartphones, geolocation systems is one of the most researched topics in computer science. The parallel can be drawn for eye-trackers. We firmly believe that with the maturing technology and miniaturization of eye-trackers, there will be an explosion of eye-tracking research and applications in the next few years.

2.3 Other modalities

Mainstream modalities such as images and audio are representations of the physical world, while textual data is the high level abstractions and interpretations of the physical world. In comparison, eye-gaze data contains information about both the bottom-up saliency of the physical world and the top-down cognitive and emotive interpretation of the visual concepts. Eye gaze patterns in response to an image is similar to taking a survey at a sub-conscious level. We ask questions by showing visual stimuli. Viewers reply by directing their attention driven by the visual stimuli. Their fixations are correlated to the bottom-up saliency and top-down influences such as tasks. Eye-gaze research complements the current multimedia processing techniques by bridging the semantic gap between the low level features and the high level abstractions.

From the perspective of Human Computer Interface (HCI), some of the common input modalities are keyboards, mices and touch-screens (Ni et al., 2014). All of these require explicit efforts by the users to interact with the systems. Eye-gaze input are implicit and have very short latency time.

Eye-gaze also compares favorably to other emerging modalities which also measure the physiological responses to visual stimuli. While there are many techniques of measuring the physiological response to visual stimuli,

we only consider those which do not require special environments such as a magnetically shielded room or those which involve invasive processes. For example, functional magnetic resonance imaging (fMRI) requires a shielded room and electrocorticography (ECoG) requires surgery to implant the electrobe grid. Thus we will compare against Electroencephalography (EEG), facial expressions analysis, human-computer interaction analysis (i.e. mouse-events) and textual responses (i.e. questionnaires and tagging). Table 2.2 shows the comparisons for spatial resolution, time to measure (latency), conscious and purposeful consideration and obtrusiveness.

With higher spatial resolution, advanced analysis techniques can be applied to infer saliency (Lang et al., 2012) and high-level mental influences. Low latency allows for real-time interactive applications such as advertising (Yadati et al., 2013) or content customization. Without the need for conscious and purposeful consideration during the measurements, problems such as dishonesty and carelessness can be minimized. It takes considerable amount of concentration to overtly control fixations while saccades are reflexive responses and uncontrollable. Non-obtrusive and non-contact devices which are both more comfortable to use and are less likely to interfere with the normal activities. Covert analysis is also possible with non-obtrusive and non-contact devices. Eye-gaze has clear advantages for these factors.

Modality	**Spatial Resolutions**	**Latency**	**Purposeful** *	**Obtrusive**
Questionnaires	Semantic regions	10 *secs*	Yes	Yes
Mouse-events	Screen resolution	1 *sec*	Yes	Yes
Touch	Screen resolution	1 *sec*	Yes	Yes
Facial expressions	Whole visual field	1 *sec*	No	No
EEG	Whole visual field	15 *msec*	No	Yes
Eye-Gaze	0.01 degree	0.5 *msec*	No	Varies **

Table 2.1: Comparison of modalities. *Purposeful refers to the conscious and purposeful decisions which *must* be made by the human subjects before the measured physiological responses can happen. **There are video-based eye-trackers which are non-obtrusive as well as head-mounted or chin-rested types, depending on the application's requirements.

2.4 Current research

There are multiple computer science areas which are using eye-tracking data. In computational human visual attention research, the eye-gaze data acts as a reliable proxy for overt attention (Borji et al., 2013a). Eye-gaze is also an emerging new biometric (Zhang et al., 2013). In user-interface studies, the eye-tracking data are used either in selective systems, gaze contingent devices or eye-gaze models for avatars. There are also studies on using eye-gaze data to solve existing computer vision or multimedia problems (Borji and Itti, 2013). Yet, there is no unified and formal framework across these different areas.

There are also other uses of eye-gaze data in neuroscience, psychology, medical diagnosis and usability studies (Duchowski, 2002; Frintrop et al., 2010). As these fields do not involve computational models, our framework will not be directly applicable to them.

2.4.1 Computational Human Visual Attention

One of the more common use of eye-gaze data is in the computational Human Visual Attention research. It aims to develop more accurate representation of the human visual attention (HVA) system. The eye-gaze data, especially the fixations, serves as an important proxy to the overt part of HVA. As such, there is an increasing number of dataset of eye-gaze data being generated by this research community (Bruce and Tsotsos, 2006; Judd et al., 2009; Ouerhani et al., 2004; Le Meur et al., 2006; Zhao and Koch, 2011).

There are two types of underlying structure, which is either based on neural networks (connectionist models) or on a collection of gray-scale maps (filter models). Regardless of the choice of structures, HVA researchers subscribe to the bottom-up/top-down factors that drive attention. The bottom-up cues refer to the characteristics of a visual scene (stimulus driven), whereas top-down cues are determined by cognitive phenomena like knowledge, expectations, reward and current goals (Frintrop et al., 2010). Bottom-up attention is fast, involuntary, and most likely feed-forward. Top-down attention is slow, task-driven, voluntary, and closed-loop (Borji and Itti, 2013). The prevailing view is that bottom-up and top-down attention are combined to direct our attentional behavior (Corbetta and Shulman, 2002). Methods are actively being researched on the integration of both types (Frintrop et al., 2010).

2.4.2 Computer Vision and Multimedia

Eye-gaze serves as an increasing popular modality for computer vision and multimedia research. It helps to bridge the gap between the low-level representation and the high-level semantics. Some examples are: segmentation (Ramanathan et al., 2010), image and video compression (Katti et al., 2011), object detection and recognition (Ramanathan et al., 2009), scene

classification (Jaimes et al., 2001), content customization (Yadati et al., 2013), surveillance video summarization (Vural and Akgul, 2009), and activity recognition (Bulling et al., 2011).

Its theoretical foundation lends heavily from the computational HVA's top-down/bottom model. Many of such applications model the top-down influences on the eye-gaze to solve some problems which are proving to be extremely difficult from pure bottom-up approaches.

2.4.3 Biometrics

Eye-gaze is an emerging behavioral biometrics (Bednarik et al., 2005; Holland and Komogortsev, 2011; Rigas et al., 2012; Kinnunen et al., 2010). Being a behavioral biometric, it offers several advantages over the mainstream biometrics such as fingerprints and face recognition. Liveness is guaranteed. Liveness detection is an important criteria in biometric systems as it prevents spoofing and replay attacks. It is also revocable. If the biometric template is stolen, it can be rendered to be invalid, thus preventing future access. Compared to face recognition, it can perform well in challenging scenarios such as identical twins identification (Zhang et al., 2013). The theoretical model is that eye-gaze is mainly dependent on the identity of the viewer and the visual stimuli.

2.4.4 Human Computer Interface

There are 3 main use-cases in Human Computer Interface (HCI) research which uses eye-gaze. Firstly, eye trackers can be used as a complement or replacement of current input devices. Besides the advantages as discussed in Chapter 2.3, eye-gaze offers an alternative for people who are unable to use conventional input devices effectively, e.g. amputees (Bednarik et al., 2012).

Secondly, the eye-gaze are used in Gaze-Contingent Displays (GCDs).

Area	Factors	Features
HVA	Bottom-up/Top-down	Fixations, Saccades
CV/MM	Bottom-up/Top-down	Fixations, Saccades, Dilations
Biometric	Identity	Micro-saccades, Scanpaths
HCI	Layouts/Objects/Tasks	Fixations, Saccades, Scanpaths, Dilations

Table 2.2: Comparison of areas of research. This table clearly illustrates that direct comparisons of research across different areas are challenging without a unifying formal framework.

Such system can tailor the display so that most informative details of the display are generated at the point of gaze but are degraded in some way on the periphery. The purpose of these displays is usually to minimize the bandwidth requirements (Duchowski, 2002).

In the third case, eye-gaze can be used to improve interactions with virtual characters or environment. Accurate modeling of eye-gaze in virtual characters is shown to improve the overall user's experience (Steptoe et al., 2009).

In all 3 cases, current HCI models assume that the eye-gaze is a function of the stimuli and the intentions of the user.

Chapter 3

VIP Framework

`Essentially, all models are wrong, but some are useful.`

`- George E. P. Box (Renowned Statistician)`

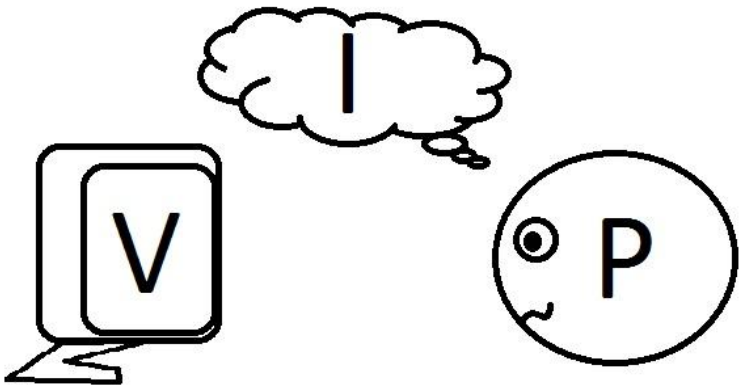

Figure 3.1: The VIP factors which will affect eye-gaze. V: visual stimuli, I: intent and P: person. All 3 factors will affect the eye-gaze of the viewer. However, in current research models, only one or two of the factors are considered.

In a controlled environment, eye-gaze information of a healthy subject is the automatic and mostly subconscious responses of the the viewer's mental processes to the stimulus. Eye-gaze information broadly refers to the raw eye-tracking data as such eye-gaze positions, blinks and dilations as well as features such as fixations, saccades, pursuit movements and scanpaths.

The visual stimulus can be an image, a video (including the audio portion), a binocular image or an interactive stimulus such as video game. Over the past 25 years, there are extensive and active studies of the properties of the stimulus which affects the eye-gaze (Borji and Itti, 2013; Frintrop et al., 2010). For example, audio features are found to be important in videos in a recent research (Song et al., 2013; Chen et al., 2014). In visual saliency literature, this is known as the bottom-up cues which include color, brightness contrast, orientation and audio.

Our research shows that there are 2 endogenous factors which affect the viewer's mental processes, namely immediate mental processes and persistent personal traits. Firstly, the immediate mental processes and conditions have strong and obvious influences (Yarbus et al., 1967; Schleicher et al., 2008; Bradley et al., 2008). These processes include the top-down influences in the visual saliency literature. The top-down influences are the knowledge, expectations, reward, and current goals (Frintrop et al., 2010). This proposal also include emotions and fatigue as examples of immediate mental factors. In other words, what the viewer wants, knows and feels at the time of viewing are the immediate mental factors which affects eye-gaze. For brevity, these factors will be coined as *intent*.

Secondly, there is recent psychological research which shows that persistent traits of the viewer can affect the eye-gaze. These traits are stable characteristics of the viewer which persist over months, years or even lifetime. Some persistent traits are the viewer's identity, gender, age and personality types etc. While acknowledged by some visual saliency researchers to be a factor which affects attention, this factor is not known to be studied by them (Borji and Itti, 2013; Judd et al., 2012). In psychology research, recent studies show that different groups of people have different gaze patterns. Goldstein et al. (2007) noted that "there are some significant differences in the observation behaviors between gender and age groups"

when watching movies. Chua et al. (2005) demonstrated that there are cultural differences in eye-movements. Personality has also been discovered as important in gaze modulation (Risko et al., 2011). Shen and Itti (2012)'s work on visual attention during listening shows that the top-down influences are modulated by gender. There are also identification systems which uses eye-gaze information as a biometric (Bednarik et al., 2005; Holland and Komogortsev, 2011; Rigas et al., 2012). These studies show that the identity, demographic factors and personality types of the viewer can affect eye-gaze. In layman's terms, the "who" and the "type" of the viewer are the persistent traits. These will be referred to as *personal* traits. We would like to emphasis that we are the first to identify and highlight the importance of personal traits in computational eye-gaze research.

3.1 Formal definitions

We then formally define the eye-gaze data, E, as follows:

$$E = g(\{t_i, x_i, y_i, p_i, q_i, d_i, s_i, b_i, c_i\}_{L/R}) \tag{3.1}$$

where g is a function of a *sequence* of eye-gaze data which

- i: the sequence number, i=1,2,...,n where n is the number of samples.
- t_i: time-stamp of the eye-gaze is related to the sampling rate. Usually the intervals are fixed.
- x_i: horizontal coordinates of the eye-gaze.
- y_i: vertical coordinates of the eye-gaze.
- p_i: horizontal location of the eye in the camera image (video-based eye-tracker only).
- q_i: vertical location of the eye in the camera image (video-based eye-tracker only).
- d_i: distance of the eye to eye-tracker.

- s_i: pupil size of the eye. (diameter or area)
- b_i: eye's opening magnitude. if $b_i = 0$, x_i,y_i and s_i are undefined since the eye is shut.
- c_i: tracking quality. (e.g. 0 = bad, 1 = excellent)
- L/R: left or right eye. Disparity can be used to compute depth or motion.

Together with other auxiliary data, such as position of eye-tracker relative to screen/object of interests, 3D position of the eye-gaze can be computed. x_i and y_i are the coordinates of the eye-gaze on the stimulus, e.g. image or video. p_i and q_i are the locations of the eyes in the video-based eye-trackers. A video-based eye-tracker is typically a infra-red camera with specialized software to map the eyes images recorded into the eye-gaze on the stimulus. Examples of g are fixations, saccades and scanpaths vectors. If g is the sequences of fixations, then each fixation u_j is defined as:

$$u_j = \{\bar{x}, \bar{y}, \bar{s}, \bar{b}, t_{start}, t_{end}\} \tag{3.2}$$

where $\bar{x}$ is the mean value of all x_i in the fixation. Similarly, $\bar{x}$,$\bar{s}$ and $\bar{b}$ are the mean value for the respective features. The t_{start} is the start time of the fixation, t_{end} is the end time of the fixation.

Similarly, each saccade v_j is defined as

$$v_j = \{x_{start}, y_{start}, t_{start}, x_{end}, y_{end}, t_{end}, \hat{v}_{peak}\} \tag{3.3}$$

where x_{start} and y_{start} are the start position of the saccade and x_{end} and y_{end} are the end position; $\hat{v}_{peak}$ is the peak speed of the saccade. A saccade is a ballistic trajectory which accelerates to its maximum speed within a very short period of time.

Then a scanpath can be defined as an ordered list of fixations and saccades:

$$w = \{u_1, v_2, u_3, u_4, v_5, ...\} \tag{3.4}$$

Since eye-gaze is influenced by the visual stimulus, intent and person; E can also be defined as a function of the 3 factors:

$$E = f(V, I, P) \tag{3.5}$$

where V is the visual stimulus' feature vector, I is the immediate mental states feature vector and P is the set of persistent personal attributes. Examples of V are the color and contrast feature vectors. Examples of I are tasks, skill levels or emotion states and emotion intensity. Examples of P are identity and gender.

In the ideal situation, g and f are equivalent. However, due to sensor's noise, computational limitations and incomplete model etc., they are not exactly the same. In the computational models which we review, the objective of the system is to minimize some application-specific error measure between the ground-truth and the system's results. Hence, "$\approx$" is taken to mean the minimization of the error measure on the both sides of the equation in this thesis.

Hence,

$$E \approx f(V, I, P) \tag{3.6}$$

We called this the VIP framework. With this framework as a reference, the features, computational model and assumptions of applications and research problems can be formally described and compared. New research directions are also easier to be discovered by identifying gaps of existing models. We will next survey the current models and applications and how they are completely defined by our framework.

3.2 Current models

Without loss of generality, consider the special case of which E depends only on V and I. Then either P is a constant or that E is independent of P. If P is a constant c, then we can rewrite $f(V, I, c)$ as $\underset{P=c}{f}(V, I)$. If P is not a constant, then f can be simplified to $f(V, I)$. For both conditions, we will refer to the simplified equations as the VI model.

3.2.1 V models

Unless otherwise stated, V models in this section assume that without an explicit goal, attention is predominantly dependent on bottom-up cues (Elazary and Itti, 2008; Judd et al., 2009). In other words, E is independent of P and I is generally assumed to be *neutral*. That is there is no active goals, knowledge or emotions which can affect eye-gaze information. Thus $\underset{I=neutral}{f}(V)$ defines these models.

The bottom-up fixation prediction algorithms are examples of applications which uses V model. The objective of these algorithms is to find some $\underset{I=neutral}{f}(V)$ such that the error measure between $\underset{I=neutral}{f}(V)$ and E is small. The error measure is usually defined as the AUC or area under the Receiving Operating Characteristic (ROC) curve (Hou et al., 2012). The main directions of research for bottom-up approach are developing better algorithms, i.e. $\underset{I=neutral}{f}$, and using new features of V. For example, Lang et al. (2012) have recently explored the influence of depth cues.

The V model is commonly used by current saliency inference algorithms as the ground-truth model. In studies whereby ground truth saliency maps were generated from gaze data as the reference for comparison against computational models, a *single average* ground-truth saliency map was generated for each image (Bruce and Tsotsos, 2006; Judd et al., 2009; Ouerhani et al., 2004; Le Meur et al., 2006; Zhao and Koch, 2011). Hence, the ref-

erence model is $\underset{I=neutral}{f^{-1}}(\{E\})$ where $\{E\}$ is the set of the fixations for all human subjects and $\underset{I=neutral}{f^{-1}}$ is a function, e.g. Gaussian filter (Judd et al., 2012), which outputs the group-truth saliency map.

Borji et al. (2013a) had performed a comprehensive comparisons of the 54 saliency models over 3 natural images datasets and 2 video datasets. The accuracy of the saliency prediction algorithms have been steadily increased over the years, owing to the better understanding of the various low-level cues which affects saliency.

The image segmentation problem is another open research problem which has successfully exploited the eye-gaze data for better accuracy (Ramanathan et al., 2010). Based on the premise that the human eye invariably fixates within the interior of an object, the algorithm attempts to find the set of boundary contours surrounding the fixation. The segmentation problem can be effectively transformed to an energy minimization problem. By using multiple fixations, its performance is better than the single random fixation method proposed by Mishra et al. (2009). Mishra et al.'s method is in turn better than pure image-based segmentation algorithms (Bagon et al., 2008; Arbeláez and Cohen, 2008). The assumption is that humans *generally* fixates on the most salient objects. The segmentation algorithm $h(f^{-1}(E), V)$ such that $f^{-1}(E)$ localizes the most salient objects in the visual stimulus.

Ramanathan et al. (2009) observed that affective concepts are consistently fixated upon by a majority of subjects. These concepts may correspond to individual objects or interactions between two objects (actions). In this application, $\underset{I=free-viewing}{f^{-1}}(E)$ localizes the regions of interests (ROI) in the visual stimuli V. E is the vector of fixations and the bi-partitioning of the exploratory saccade movements. The ROI are then fused with the content analysis of text and visual stimulus to localize and label affective objects in images. The multi-modal framework compares favorably against

Viola-Jones face detector which uses only image-based analysis. As the stimuli for their experiments are affective, $I = free - viewing$ instead of *neutral*.

The real-time surveillance video summarization system proposed by Vural and Akgul (2009) can mix actions from different frames into the same video for more compact videos. A real-time automated algorithm will detect video sections which actions has occurred. Filtering is performed on the detected video section based on the fixations of the human operator. The $\underset{I=surveillance}{f^{-1}}(E)$ computes the ROI within the video frames V. For this application, I is implicitly assumed to be mental state of a security personnel at work which is termed *surveillance*. The *task* of general surveillance, e.g. looking for suspicious actions and *domain knowledge* such as familiarity about the monitor environment are expected to be part of *surveillance* state of mind.

Avatar's eye-gaze modeling is one important factor for an immersive experience in a virtual environment. Steptoe et al. (2009) implemented a eye-gaze model to compare against tracked gaze and static gaze. Their eye-gaze model determines the field-of-view (FOV) of the avatar and randomly distributed its fixations and saccades on the faces of avatars and objects within the FOV. Hence, $E \approx f(V)$.

Generally, the V applications improve upon image-based algorithms by integrating E into the algorithms. There are many other such problems which benefited from the eye-gaze information.

3.2.2 I models

The *general* I models assume that some I can completely determine the specially selected E, i.e. $E \approx f(I)$. One such example is the activity recognition system by Bulling et al. (2011). They recorded saccades, fixations and blinks using an wearable EOG. It can classify 5 activity classes:

copying a text, reading a printed paper, taking handwritten notes, watching a video, and browsing the Web. It opens up the wider applicability of eye-gaze data to other activities that are difficult, or even impossible, to detect using common sensing modalities. $f^{-1}(E)$ identifies the activities I from the eye-gaze features E. In this paper, while the experiment was setup in an office, the system did not limit its applicablity to this specific environment.

Another example is the "Midas-touch" problem in gaze-based interactions systems. The problem is to infer I from E so that the systems can determine if a fixation is observing or actioning (e.g. issuing a command). Bednarik et al. (2012) have used the features extracted from fixations, saccades and pupillary responses to determine the intentions of the user. Their experimental results indicated that fixations and saccades features are more reliable than pupillary responses for predicting intentions.

Asteriadis et al. (2009)'s system uses a fuzzy inference system to estimate learner's attention states, I (Attentive/Inattentive) from eye-gaze E. A Sugeno-type fuzzy inference system was used (Takagi and Sugeno, 1985).

3.2.3 P models

As eye movements are counterfeit resistant due to the complex neurological interactions and the extraocular muscle properties involved in their generation, they have been proposed as a viable biometric by various papers (Holland and Komogortsev, 2011; Rigas et al., 2012). In these papers, the stimulus and the tasks are the same during the training and testing phases. Hence, $P \approx \underset{V=c1,I=c2}{f^{-1}}(E)$.

Kinnunen et al. (2010) implement a stimulus V independent eye-gaze biometric. They identified the histogram of all angles the eye gaze travels during a short period, few seconds, as a potential predictor of a person's identity regardless of V and I. Hence, $f^{-1}(E)$ infers the person's identity

P from the histogram E. Their methods are unlike those which use the same task and stimulus for identification.

Our proposed system uses the features extracted from videos of subjects talking to distinguish identical twins (Zhang et al., 2013). Out of the 6 features, 3 of them are gaze data: gaze change, pupil movement and eye open magnitude. In our system, the V varies from the bedroom, recording studios to convention halls filled with people. Therefore, the V is not of consequences to the accuracy. The groups of people P ranges across different age, gender, ethnicity etc. One interesting point is that our system is able to distinguish between identical twins.

To the best of our knowledge, other than biometric applications, there are no application which infers P from E. However, there are many advantages of using eye-gaze to infer other personal attributes as compared to traditional medium such as questionnaires and vision-based approach. Thus, we propose a novel implicit trait inference problem which is to accurately infer the personal traits from the eye-gaze. It assumes that given the same stimulus, viewers having similar immediate mental states (*free-viewing*) but differing personal traits will have different eye-gaze patterns. Hence, $\underset{V=c,I=free-viewing}{f^{-1}}(E)$ will infer P. We achieve the accuracy of 0.92 with 52 subjects viewing 2 images for introvert/extrovert classification. Further details are presented in Chapter 5.

3.2.4 VI models

This model assumes that eye-gaze is dependent on both the visual stimulus features and the immediate mental states of the viewer, e.g. tasks or emotions.

The fixation prediction algorithms which combines both top-down influences and bottom-up cues are examples of applications which uses VI model. The objective of these algorithms is to find some $h(V, I)$ such that

the error measure between $h(V, I)$ and E is small. The main directions of research are the methods of combining the features V and I and using new features (Frintrop et al., 2010).

In implicit tagging applications, affectiveness of the stimulus is automatically assessed from the viewer's various physiological signals, including pupillary dilation (PD) (Pantic and Vinciarelli, 2009). The PD is known to be influenced by emotions (I) and light intensity (V). Gao et al. (2009) attempted to use Adaptive Interference (AIC), with H* time-varying adaptive (HITV) algorithm to determine the emotions of the viewer. Hence, $I \approx f^{-1}(E, V)$.

Katti et al. (2011) performed personalized affective video summarization by using pupillary dilation (PD) to detect important and affective segments of the videos. $f^{-1}(E, V)$ is the importance and affective of the video segments. For video summarization, PD infers the affective of the video segments and shot boundaries in the video discards the frames with peak PD.

Yadati et al. (2013) have proposed a novel method for interactive personalized advertisement insertion for a single user. The proposed system fuses in real time, the emotion type (from facial expressions), emotion intensity (from PD) and the affective values (from affective analysis of the video). The most effective advertisements are then inserted accordingly. It has better brand recall rate than the referenced affect agnostic method. $f^{-1}(E, V)$ infers the I (emotional intensity) from both E (PD) and V (image-based affective analysis of the video segment).

Samsung Galaxy S IV is a smartphone with eye-tracking capability (Samsung, 2013). It can detect whether the user is looking at the screen and adjusts its response according to the displayed task. The "Smart Stay" feature will turn off the screen if the eyes are not detected; the "Smart Pause" feature will pause a playing video if the user looks away. Thus,

$I \approx f^{-1}(E, V)$.

3.2.5 Other cases

The VP and IP models are not sufficiently explored by researchers. The VP model assumes that given some constant I, the eye-gaze are dependent on both the visual stimulus and the personal traits. One example application which we will show in the Chapter 6 is the trait-specific fixation predictors. With the same I, E (fixations) can be *more accurately* predicted for some *specific* images using an trait-specific saliency map, e.g. male saliency map to predict male's fixations, compare to the V predictors.

The IP model assumes that E is either independent of V or that V is fixed. We do not know of any application or research problem with such assumptions. One potential research problem would be the co-inference of I and P from E that is $(I, P) = f_V^{-1}(E)$. For example, given some specially selected video and the eye-gaze features, the algorithm can infer gender *and* emotions.

From our extensive literature survey, there is no research problem which is formulated as the most general VIP model. This is clearly a big and interesting gap to be filled. One of the first step is to build a dataset which consists of all 3 factors. Much scientific insights can be gained from a comprehensive dataset which consists of all 3 factors. For example, new discoveries about the relationship and patterns can be found. Co-inference of 2 or even 3 factors may be possible. We propose a novel application of implicit just-in-time profiling in Chapter 7. It is the first system to incorporate the 3 factors.

Table 3.1 summarizes the features and applications for the various VIP models. From the table, it is clear that V models are well-researched and there are research gaps to be filled in the other models, especially the various combinations of P.

Model	Features	Applications/problems	References
V	color,	fixation prediction,	Borji and Itti (2013); Lang et al. (2012)
	brightness,	bottom-up saliency,	Judd et al. (2009); Zhao and Koch (2011)
	contrast, depth	image segmentation,	Ouerhani et al. (2004); Le Meur et al. (2006)
	region of interests	image annotations	Mishra et al. (2009); Ramanathan et al. (2009)
	contrast, depth	image segmentation,	Bruce and Tsotsos (2006); Ramanathan et al. (2010)
I	tasks,	activity classification,	Yarbus et al. (1967); Bulling et al. (2011)
	fatigue,	fatigue detection,	Bednarik et al. (2012); Asteriadis et al. (2009)
	emotions	emotions classification	
P	identity,	biometric,	Zhang et al. (2013); Kinnunen et al. (2010)
	personality		Holland and Komogortsev (2011); Rigas et al. (2012)
	demographic	*traits inference*	*Chapter 5*
VI		saliency models,	Frintrop et al. (2010); Katti et al. (2011)
	V and I	video summarization, interactive advertisement	Yadati et al. (2013)
VP	V and P	*trait-specific fixation prediction*	*Chapter 6*
IP	I and P	$(I, P) \approx f^{-1}(E)$	Open area
VIP	V, I and P	*VIP dataset*, *VVIP dataset* `Eye-2-I`	*Chapter 4* *Chapter 7*

Table 3.1: Comparison of various VIP models. The 3 applications: personal trait inference, trait-specific fixation prediction and `Eye-2-I` system are our contributions in this proposal. The VIP and VVIP datasets are described in the Chapter 4. We also suggest open research problems for IP model.

Chapter 4

Datasets

`It is a capital mistake to theorize before one has data.`

`- Sherlock Holmes, A Study in Scarlett (Arthur Conan Doyle)`

The importance of publicly available datasets cannot be overstated, especially for computer science research. Comparison of new methods against existing ones will be more systematic and reliable with standard benchmarks. Without a common dataset, it is impossible for researchers to reproduce the empirical experimental results.

Publicly available eye-tracking datasets are fewer than other modalities such as images, videos and audios. Some of the more popular and important eye-gaze datasets are cited in Winkler and Subramanian (2013)'s paper and the MIT website: `http://people.csail.mit.edu/tjudd/SaliencyBenchmark/`. According to Winkler and Subramanian (2013), there are over two dozens eye-gaze datasets for images/videos in public domain.

Eye-gaze datasets are costly in manpower, time and money to acquire. Eye-gaze datasets takes several months to collect. Firstly, there is inherent logistical and administrative challenges in any user study. For example, we needed to get approval from the NUS Institutional Review Board before

we can commerce our experiments. The approval process itself can take a month or more. Any amendment in the protocol had to be submitted and approved. The recruitment of subjects is time-consuming. Secondly, prices of accurate eye-trackers are higher than other sensors, such as video cameras or GPS. This limits the collection of eye-tracking dataset to researchers with sufficient financial resources or existing equipment (Pal et al., 2009). Due to budget constraints, we had access to only one eye-tracker. As such, we can only have one subject per session. Thirdly, as very few people have experience with eye-trackers, a fully self-help experimental setup is not possible. An experimenter had to present to provide assistance, especially for the calibration process (Chapter 2.1). Therefore, crowd-sourcing methods cannot be used to collect eye-tracking data

To the best of our knowledge, we are the first to collect and publish eye-tracking datasets coupled with comprehensive viewer's demographic and personality traits. From this thesis, it is clear that personal traits are important factors affecting eye-gaze. Our VIP framework will promote the recording and release of the personal traits of the subjects as a standard practice for eye-gaze datasets.

Our VIP dataset can be downloaded from `http://mmas.comp.nus.edu.sg/VIP.html`. The VVIP dataset will be made available at the same URL after our submission is accepted.

4.1 VIP Dataset

The VIP dataset is the first publicly available dataset with complete VIP factors. It will propel computational eye-gaze research into unexplored territories. The dataset consists of 2 tasks: free-viewing and anomaly detection; and it contains neutral/affective images (I factors).

4.1.1 Data collection protocol

75 participants were recruited from a mixture of undergraduate, postgraduate and working population. The male and female subjects are recruited separately to ensure an even distribution. Participants have corrected-to-normal or normal eyesight. If spectacles are worn, the lenses are NOT more than 350 degrees. There is no restriction for contact lens. The reason is due the capability of the eye-tracker.

They were tasked to view 150 images. They were either instructed to free-view (i.e. without assigned task) or to perform anomaly detection. Each image was displayed for 5 seconds, followed by 2 seconds viewing of a gray screen. The images were displayed in random order. Their eye-gaze data was recorded with a binocular infra-red based remote eye-tracking device SMI RED 250. The recording was done at 120Hz. The subjects were seated at 50 centimeters away from a 22 inch LCD monitor with 1680x1050 resolution.

We considered carefully of the trade-off between having more accurate and clean eye-tracking data using physical restrains, e.g. chin-rest; and subjects in a more realistic setup with freedom of eye, head and body movements. As our objective is to profile the subjects implicitly and unobtrusively, the subjects were not restrained by any physical contraption, e.g. chin rest or head rest. This setup is different from most other datasets (Winkler and Subramanian, 2013). To obtain good quality eye-tracking data, the subjects were instructed to keep their eyes on the screen during the presentation of the visual stimuli and to minimize movements. We noted that some subjects did not follow these instructions. Some subjects feedbacked about fatigue. These data are not rejected and are published together with others. Overall, the eye-tracker performance was satisfactory even without to restrain the subjects' head movements.

Chin-rest will definitely improve reliability and accuracy. However,

chin-rest was not used because the remote SMI RED250 eye tracker is designed to work with head movements, so chin-rest is optional. The recording rate is set to 120Hz which is well within the maximum 250 Hz of the device.

The recorded eye-gaze data were preprocessed by the vendor's software to extract the fixations.

4.1.2 V features

The images were selected from the NUSEF (Ramanathan et al., 2010) dataset, which contains both neutral and affective images is selected. Out of 758 images, 150 were randomly selected. The set of selected images contains a wide variety of semantic and affective content. There are faces, portraits, groups of people, animals, objects and scenery (indoor/outdoor). The set of images consists mostly of natural images, with a few paintings and computer generated images. We made the difficult decision to exclude nude images in our VIP dataset so that we can recruit subjects from a more diverse population, e.g. religious people.

Some of the images are also expected to evoke emotions such as fear, disgust and joy in some people. There are also some religious, cultural and political images.

The choice of sampling from NUSEF images over other datasets is due to its balanced mixture of affective and non-affective content. As our work is to validate that personal traits can have measurable influences over eye-tracking data (and visual attention), we immediately exclude those datasets with only simple scenes. The low level V features, such as colors and contrast, are likely to outweigh any more subtle influences from the personal traits for this category of visual stimuli. This is also known as the *pop-out* effect (Frintrop et al., 2010). We then further consider the datasets with more complex scenes, but containing mostly affective-neutral content. For

Features	References
Colors	Bruce and Tsotsos (2006); Ouerhani et al. (2004); Judd et al. (2009)
Orientation	Ouerhani et al. (2004); Judd et al. (2009)
Intensity	Le Meur et al. (2006); Judd et al. (2009)
Faces	Steptoe et al. (2009); Judd et al. (2009)
Semantic Contents	Shen and Zhao (2014)

Table 4.1: Some V features used by researchers.

these datasets, the agreement between subject's fixations are quite high. For example, in Judd et al. (2012)'s experiments, the mean area under ROC curve (AUC) is 0.922 for human subjects. While the visual stimuli may be more complex than simple scenes, the contents are mostly neutrally affective. Hence, influence of the P factors were outweighed by V and I factors for these datasets.

Our analysis of the NUSEF dataset painted a different picture. The inter-subject agreement for many images, especially affective ones such as nudes and portraits, are quite low. This is in contrast to most saliency studies which conclude that human subjects have very high agreement rates. Our choice of NUSEF images is independently validated by Borji et al. (2013b). Their experimental results show low prediction accuracy for existing computational models over emotional stimuli from the NUSEF dataset. These results can be explained by the fact that the current models did not account for personal traits. For example, the portrait images evoked different fixation patterns for male (fixated at the mouth/nose) and female subjects (fixated at the eyes).

4.1.3 I features

The subjects were either free-viewing which they were instructed to "look at the images as you would normally do"; or to perform anomaly detection which they were instructed to ''look for any unusual objects or events in the images". These tasks are designed to be applicable to all visual stimulus.

Trait	Our Proxy Metrics	References
Gender	Gender	Goldstein et al. (2007)
Age	Age Groups	Goldstein et al. (2007)
Culture	Ethnicity, Country of Birth, Nationality	Chua et al. (2005)
Religion	Religiosity	Bressan et al. (2008)
Socioeconomic	Education Level, Income Group, Expenditure Group, Specialization	Hunziker (1970)
Personality	Personality Types	Risko et al. (2011); Wu et al. (2013)

Table 4.2: Psychology studies on the correlation between eye-movements and personal traits.

70 subjects were free-viewing while 5 were detecting anomalies. There are no explicit unusual objects/things in the dataset. The purpose of the instruction is investigate the influence of different I (free-viewing vs anomaly detection). The anomaly task was chosen as it is a general task, applicable to any image. This is similar to the seminal Yarbus et al. (1967)' study.

As a future work, more data for the "anomaly detection" task will be collected.

4.1.4 P features

The subjects also provided their demographic data: gender, age-group, ethnicity, religion, field of study/work (specialization), highest education qualifications, income group, expenditure group, place of birth and nationality for the experiment. Each of these are demographic traits routinely collected by various organizations for marketing, personnel screening and advertising purposes. Some of these traits were also found to correlate with eye-gaze (Table 4.2). The histograms of the demographic traits are shown in Figure 4.1 to 4.9.

The subjects were also tasked to answer 3 questions about their personality. The personality type questions are based on the Jung's Psycho-

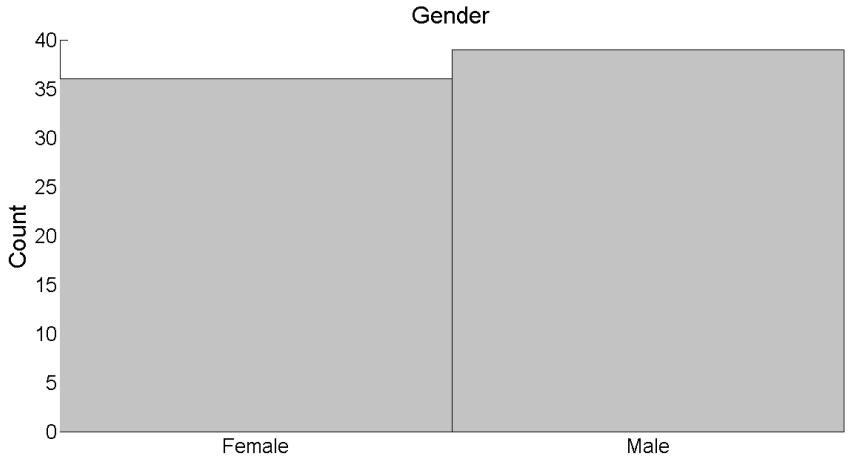

Figure 4.1: Histogram of the subject's gender distribution.

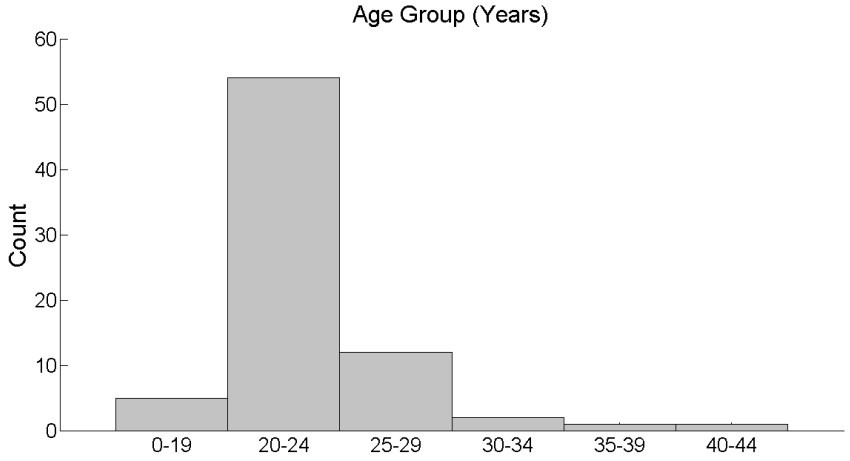

Figure 4.2: Histogram of the subject's age distribution.

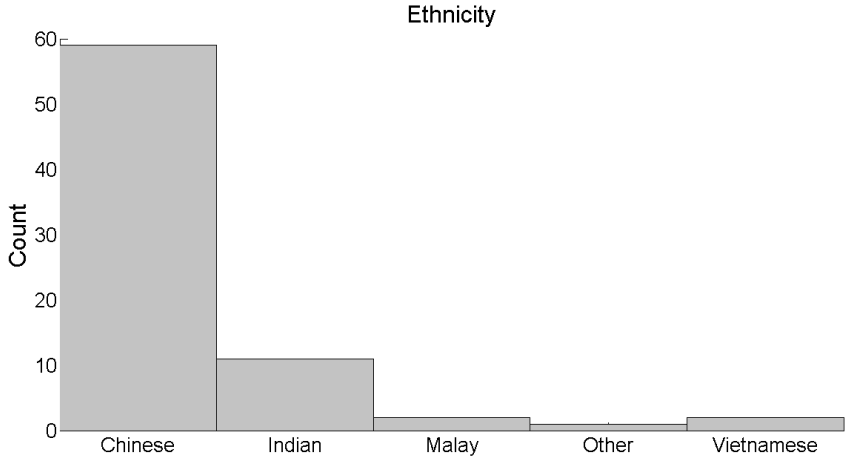

Figure 4.3: Histogram of the subject's ethnicity distribution.

logical types (Jung et al., 1991). There are only 2 possible responses for the questions. The Jung's types are very similar to the Myers-Briggs Type

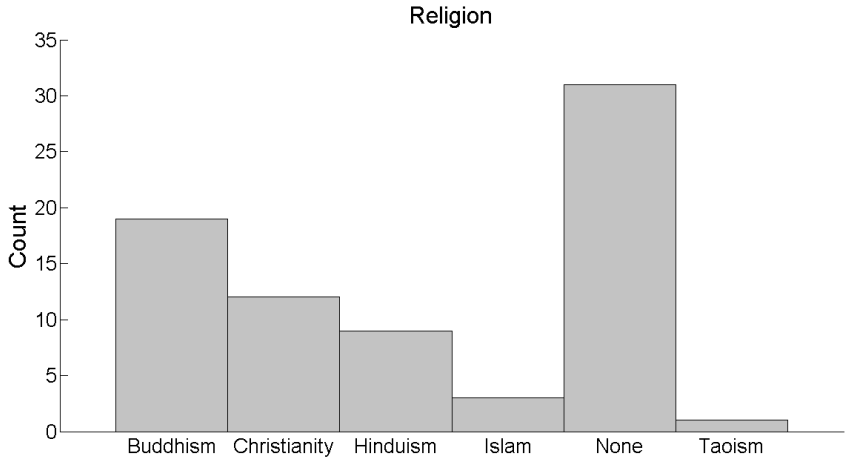

Figure 4.4: Histogram of the subject's religion distribution.

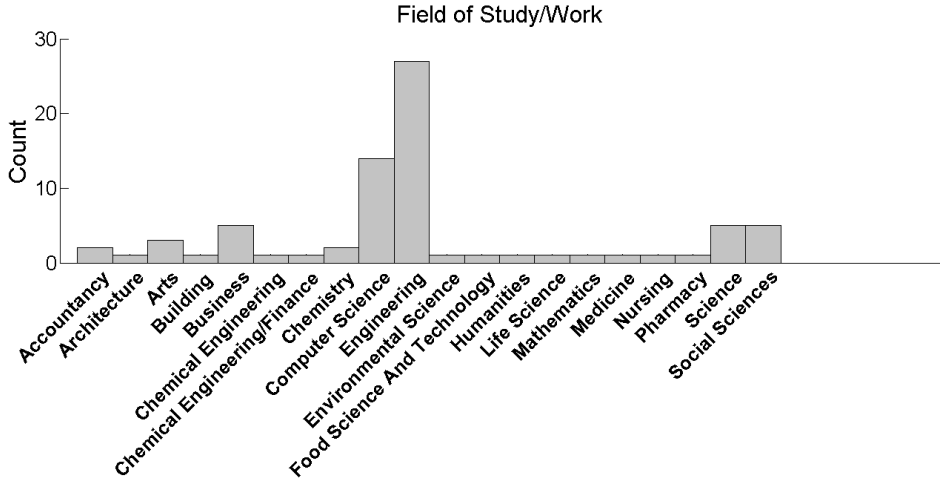

Figure 4.5: Histogram of the subject's field of study/work distribution.

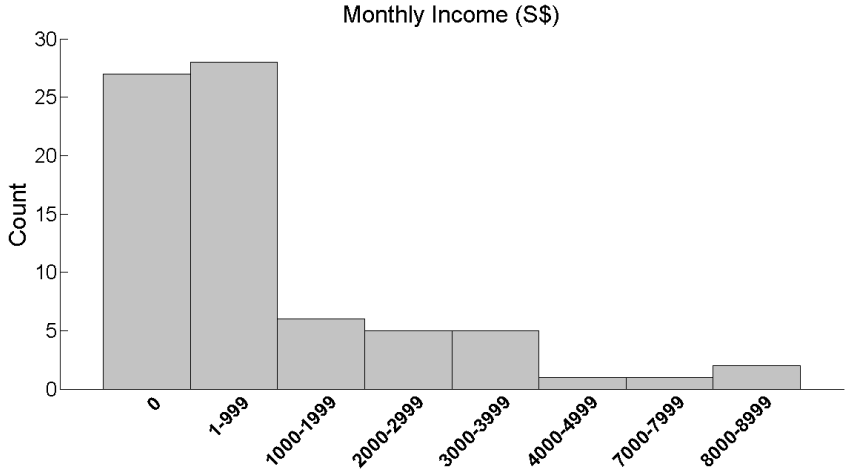

Figure 4.6: Histogram of the subject's income distribution.

Indicator (MBTI) which is widely used by the industry for personnel screening (Briggs and Myers, 1980). The histograms of the personality types are

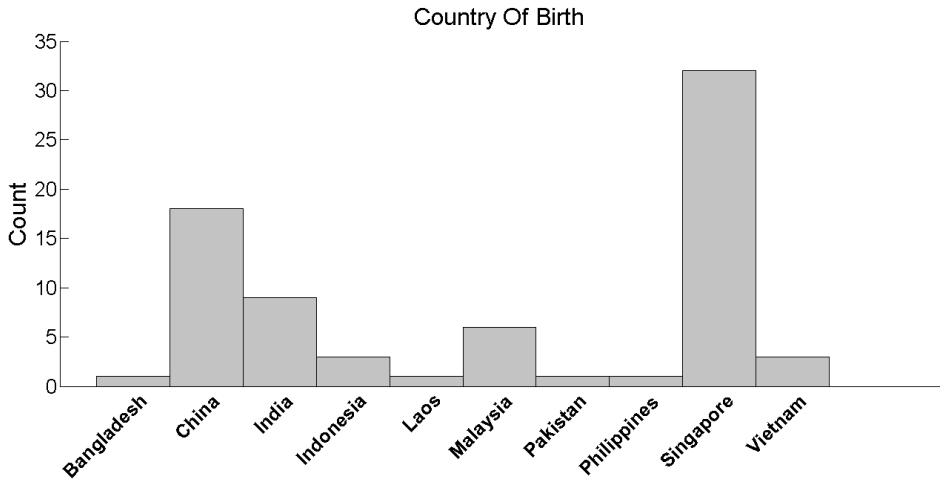

Figure 4.7: Histogram of the subject's country of birth distribution.

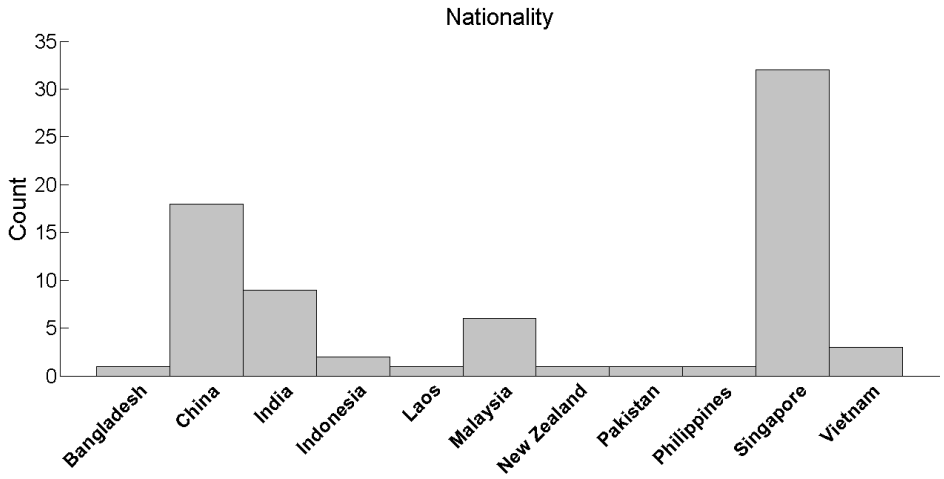

Figure 4.8: Histogram of the subject's nationality distribution.

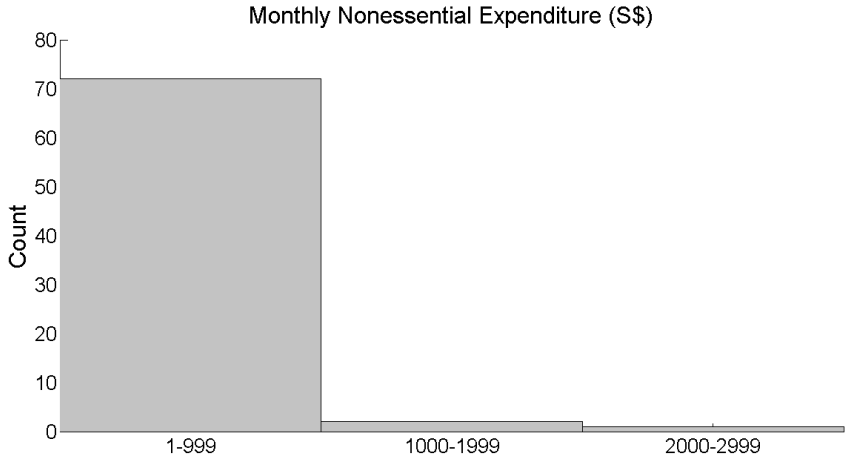

Figure 4.9: Histogram of the subject's non-essential expenditure distribution.

shown in Figure 4.10, 4.11 and 4.12. The questionnaire is presented in Appendix A.

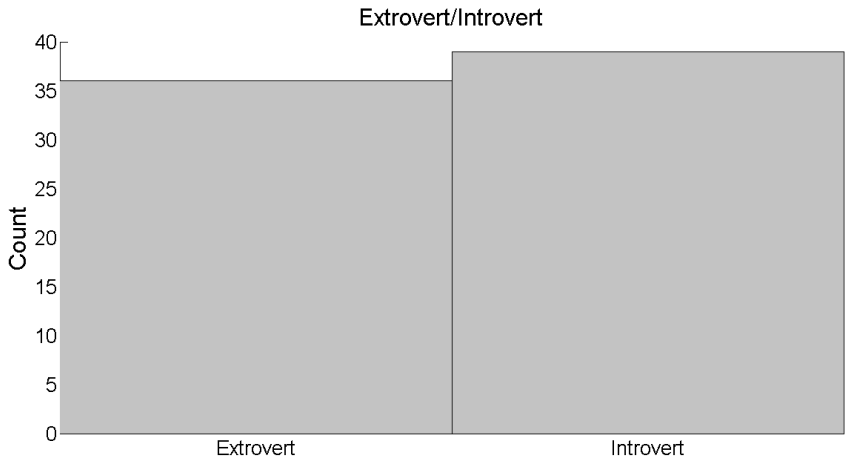

Figure 4.10: Histogram of the subject's extrovert/introvert distribution.

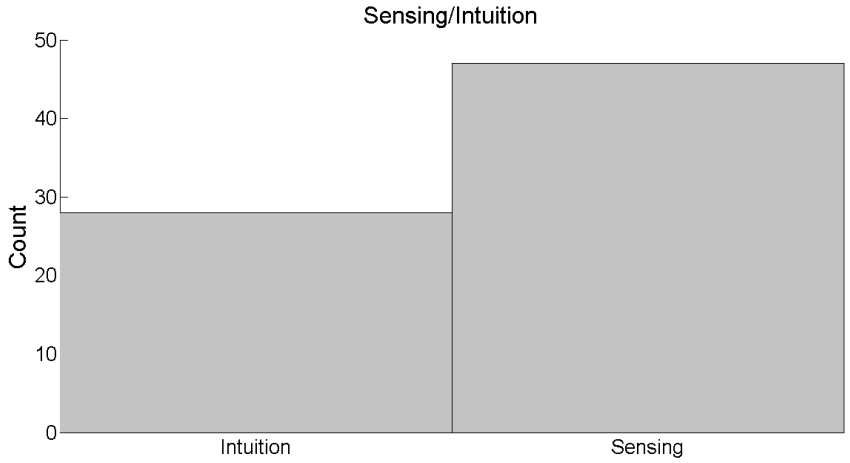

Figure 4.11: Histogram of the subject's sensing/intuition distribution.

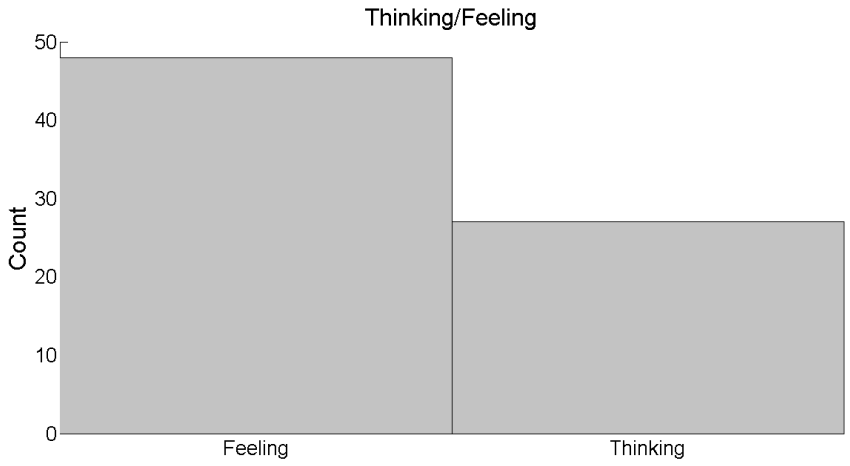

Figure 4.12: Histogram of the subject's thinking/feeling distribution.

4.2 VVIP dataset

The Video-VIP (VVIP) dataset is the first multi-modal dataset (facial expression, eye-gaze, video affects analysis and text) coupled with anonymous demographic profiles, personality traits and topics of interest of 51 participants for non-commercial and not-for-profit purposes.

4.2.1 Data collection protocol

Fifty-one participants were recruited for the paid experiment from an undergraduate, postgraduate and working adults population. Participants have corrected-to-normal or normal eyesight. If spectacles are worn, the lenses are NOT more than 350 degrees. There is no restriction for contact lens.

They were tasked to view all 4 videos in a free-viewing settings (i.e. without assigned task). Specifically, they were instructed to view the videos as they would watch in their leisure time on their computer or television.

Their eye-gaze data was recorded with a binocular infra-red based remote eye-tracking device SMI RED 250. The recording was done at 60Hz. The subjects were seated at 50 centimeters away from a 22 inch LCD monitor with 1680x1050 resolution. A web-camera is also set up to record

their facial expressions, which is analyzed by an emotion analyzer, eMotion (Gevers, 2014).

Similar to the VIP setup, there was no physical restrains. The subjects were also given instructions to keep their eyes on the screen and to remain in a relaxed and natural posture, with minimal movements. Again, we noted that some subjects did not follow the instructions. The subjects were too engaged with the content that they moved unconsciously. For example, a few subjects were laughing heartily with significant head and body movements while watching the comedy videos. The data collected are of higher quality than the VIP, due to higher engagement in content; calibration before start of each video; and more control of the calibration process. The recording rate was also lowered to 60Hz.

Figure 4.13 shows the experimentation setup in our laboratory.

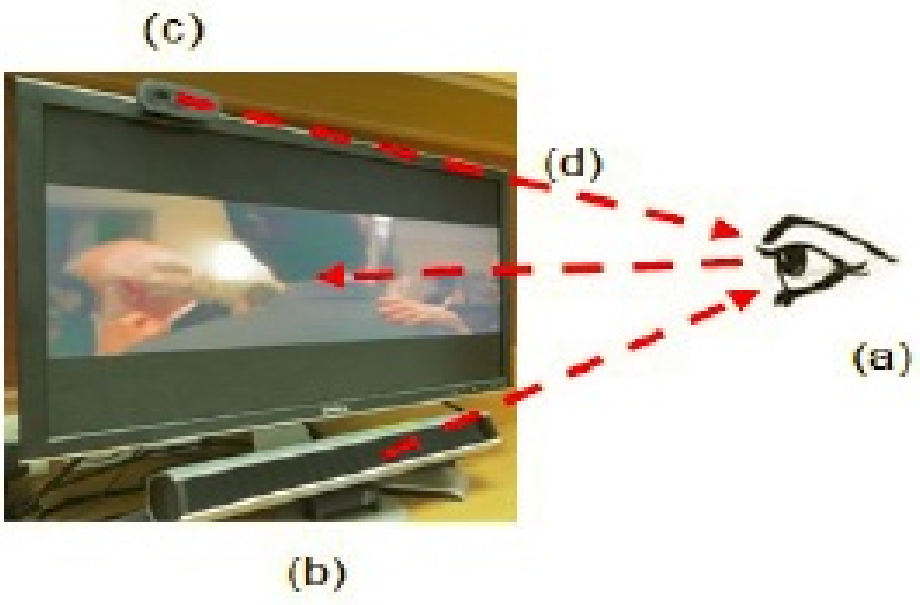

Figure 4.13: Experimental setup. (a) The user; (b) Eye-tracker; (c) Camera; (d) Stimulus

Video	Title	YouTube Category	Descriptions
Documentary	Attenborough: Beaver Lodge Construction Squad	pets & animals	The American beaver's ability to nibble wood demonstrates the stunning adaptability of these amazing mammals. In addition to creating their own lake, this family of beavers construct a make-shift fridge and wintertime snug.
Animation	Big Buck Bunny animation	film & animation	An animated short film about a Big Buck Bunny
Satire	Bloopers: Sh*t We Watch on Television	comedy	We aim to create more distinctly Singaporean videos to make you laugh and cry
Romance	Somewhere Like This	film & animation	Across clear skies. Among glowing stars. That's where I'll find us. Have you ever found yourself revisiting a chapter of your life when you least expected to? Does the story continue where it last left off? Or does it start over with new characters and conflicts? "Somewhere Like This" explores the effects of distance and time on a young couple, Scott and Irene.

Table 4.3: YouTube meta-data of the videos.

Video	Acts	Languages	Cast	Affect	Time (s)
Documentary	1	British English	1 man	Neutral, Calm	586
Animation	3	no speech	4 animals	Cheerful, Excited	596
Satire	multiple	multilingual	multiple persons	Cheerful, Excited	625
Romance	1	American English	1 man, 1 woman	Sad, Neutral	667

Table 4.4: Summary of the characteristics of the videos.

4.2.2 V features

Some videos are more likely than others to elicit eye-gaze behaviors which are suitable profiling of the different attributes. We have carefully selected 4 videos with different categories, number of acts, languages, cast make-up and affect. Table 4.3 shows the meta-data of the videos from YouTube. The characteristics of the videos are summarized in Table 4.4. The duration of each video was about 10 minutes. All videos were presented to every participants in random order.

With reference to the circumplex model of affect (Russell, 1980), the dataset contains emotional valence (type: cheerful, sad) and arousal (intensity: excited, calm) values which are computed from the visual and audio features with the algorithm as described by Hanjalic and Xu (2005). The arousal features are motion activity, shot change frequency and average energy in the audio stream. The valence features are HSV color histogram

	like	valence	arousal	Prior viewers
documentary	3.96(0.96)	6.24(1.26)	5.35(1.93)	0
animation	4.04(0.89)	6.65(1.73)	6.20(1.50)	2
satire	3.73(1.10)	7.02(1.39)	6.00(1.87)	2
romance	4.08(0.74)	3.84(1.62)	5.73(1.42)	3

Table 4.5: Summary of the subject's feedbacks of the videos. The first number is the mean and the number is parentheses is the standard deviation. *Romance* video has the highest rating and lowest valence. *Animation* video has the highest arousal. *Documentary* video has the lowest arousal. *Satire* video has the lowest rating and the highest valence. The last column shows the number of subjects who had already viewed the videos before the user study.

arts & humanities	automotive	business
finance & insurance	entertainment	Internet
computer & electronics	real estate	local
reference & education	recreation	science
news & current events	telecomms	sports
beauty & personal care	animals	games
food & drink	industries	shopping
photos & videos	lifestyle	travel
home & gardening	social network	society

Table 4.6: Topics of interest from Google Ads. The topics will be referenced by their first words (underlined) in this thesis.

and pitch from the audio signal.

4.2.3 I features

The participants are also tasked to answer questions after watching the video: rating (1-5, dislike to like), emotional valency (1-9, sad to cheerful), emotional arousal (1-9, calm to excited). They also select topics which are related to video from a list, see Table 4.6 for the listing of the topics. The participants were also asked if they had viewed the videos before. Table 4.5 shows the mean and standard deviations of the feedback for the videos. Only very few subjects have viewed the videos before the experiment.

The data also contains the normalized pupil dilation values which are indicative of emotional arousal (Bradley et al., 2008). The dataset also con-

sists of the facial expression results computed from the web-camera. The eMotion emotion analyzer tracks the face and returns a streaming probability for neutral, happy, surprise, anger, sad, fear, and disgust (Gevers, 2014).

4.2.4 P features

The subjects also answered questions on their demography and personality. The questions are identical to the VIP dataset, except that subjects also report their household income.

Chapter 5

Personal Traits Inference

While the male eye zooms in on a particular element
to the exclusion of all else, a woman's gaze flickers
from one tedious task to the next, to the point where
we can't distinguish between the importance of
mopping the kitchen floor and achieving world peace.

- Mariella Frostrup (Journalist)

We are the first to propose that personal traits (P) can be inferred from eye-gaze information (E). We define V to be constant for the training and the inferencing. I is assumed to be free-viewing. Thus, the problem is defined as:

$$P \approx \underset{V=c,I=free-view}{f^{-1}}(E)$$

where P is the persistent personal trait, e.g. gender. $\underset{V=c,I=free-view}{f^{-1}}$ is the classifier which was trained on eye-gaze information of other subjects when free-viewing the same stimulus. The information is labeled with their corresponding P. E is the eye-gaze information of the test subject. This problem is a P model since the V and I are constants.

Many of the personal traits, such as gender, age, culture and personality types are routinely collected by many organizations. These traits are

collectively known as demographic/personality profile. Companies can predict the buying behavior of customers based on their profile. Employers routinely demand prospective recruits to take personality tests. The advantages of eye-gaze over other modalities are listed in Table 2.2, that is low latency, no purposeful thoughts required and non-obtrusive. Hence, it can be deployed in situations which require real-timeliness, implicitness, discretion and/or in adversarial environment. Current profiling techniques are impossible or impractical for such situations.

Personal traits inference is analogous to taking a survey. The eye-gaze information in response to an image is similar to taking a survey at a subconscious level. Instead of questions, visual stimuli are presented. The viewers respond by directing their attention driven by the visual stimuli. Similar to the question in a survey, only eye-gaze data of selected stimulus can determine the correct value of the intended trait. Conversely, using the eye-gaze from the wrong images to infer the correct value of the intended trait is akin to asking the question, “What is your favorite color?" to know the age of an subject.

5.1 Experimental setup

We used a subset of the VIP dataset. We only used the fixation data which are collected during the free-viewing task. From these, the recorded data of the 52 subjects, 27 females and 25 males, who have fixations for more than 100 images were used for further analysis. The number of subjects are comparable to similar studies in eye-gaze experiments (Winkler and Subramanian, 2013). Only eye-gaze data from the preferred eye as chosen by the subjects were used.

From the collected trait data, we selected 5 traits which are naturally group into 2 classes and have relatively well-balanced distributions: gender,

religiosity and the 3 personality types for our experiments. For religiosity, the subjects indicate their religion in the questionnaire. If they input as "none", "atheism" or "free-thinker", they are grouped as $non - religious$. Otherwise, they are grouped as $religious$.

5.2 Features selection

As this is the first work on using eye-movement data to classify demographic and personality traits, there is no prior research to directly leverage on. From our preliminary inspections of the fixation data, we found that female has greater variations of fixation locations for some images. Miyahira et al. (2001) have found that the genders have different mean scanning time in viewing of simple drawings. The difference in emotion processing between the different personality types may also be revealed by the pupillary dilations. We thus select the potential 19 features as follows:

- mean value of the coordinates, x, y, of the fixations: $\bar{x}, \bar{y}$
- mean value of the fixations' duration: $\bar{d}$
- triangle matrix of covariance of x and y: σ_x, σ_y and σ_{xy}
- standard deviation of duration: σ_d
- normalized pupil dilation: $\hat{p} = \sigma_p / \bar{p}$
- 1^{st} fixation: x_1, y_1, d_1
- 2^{nd} fixation: x_2, y_2, d_2
- fixation with the longest duration: x_L, y_L, d_L
- total fixation duration: D
- number of fixations: N

To select the relevant features for classification, a correlation analysis method, $corrcoef$, was applied. The analysis is performed for each image separately. As an example, for the image "dog.jpg', the analysis is applied to $\bar{x}$ of all subjects and the corresponding trait of the subjects (female=1,

male=0).

The matrix $R = corrcoef(A, B)$ is related to the covariance matrix $C = cov([A, B])$ by $R(i, j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}$, where A is the feature, e.g. $\bar{x}$ and B is the trait. The zeroth lag of the normalized covariance function is used to compute the correlation coefficients and the hypothesis of no correlation. Each p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero. The correlation is defined as *significant* if $p - value < 0.05$. This means that the probability of observing a correlation due to statistical fluke is only 5%.

We want to select the features which are highly correlated with the traits' values and have low $p - value$ for many images. Since each pair (feature,image) has a 0.05 probability of being *significantly* correlated due to random coincidences, the number of expected correlated images for a feature is $0.05 * 150 = 7.5$ for the set of 150 images. Therefore, only features which *significantly* correlates with the trait for more than 7.5 images are selected. The results are summarized in Table 5.1.

Hence, the features E selected are:

- *Male/Female*: σ_x, σ_y, $\hat{p}$
- *Religious/None*: $\bar{x}$, $\hat{p}$, x_1, x_2, d_2, x_L, D, N
- *Extrovert/Introvert*: σ_{xy}
- *Sensing/Intuition*: $\bar{d}$, σ_d, y_1, y_2, d_2, d_L, D
- *Thinking/Feeling*: $\hat{p}$, y_2, d_2, x_L, y_L

The correlation analysis results shows that *Male/Female* have different variations of fixations and that their pupillary dilations are different. One example of the features differences (σ_x, σ_y) are shown in Figure 5.1.

For religiosity, the 2 groups tends to fixate on different parts of the image as shown in the $\bar{x}$, x_1, x_2. The fixation durations also differs. We noted that D and N are correlated with religiosity for many images. However,

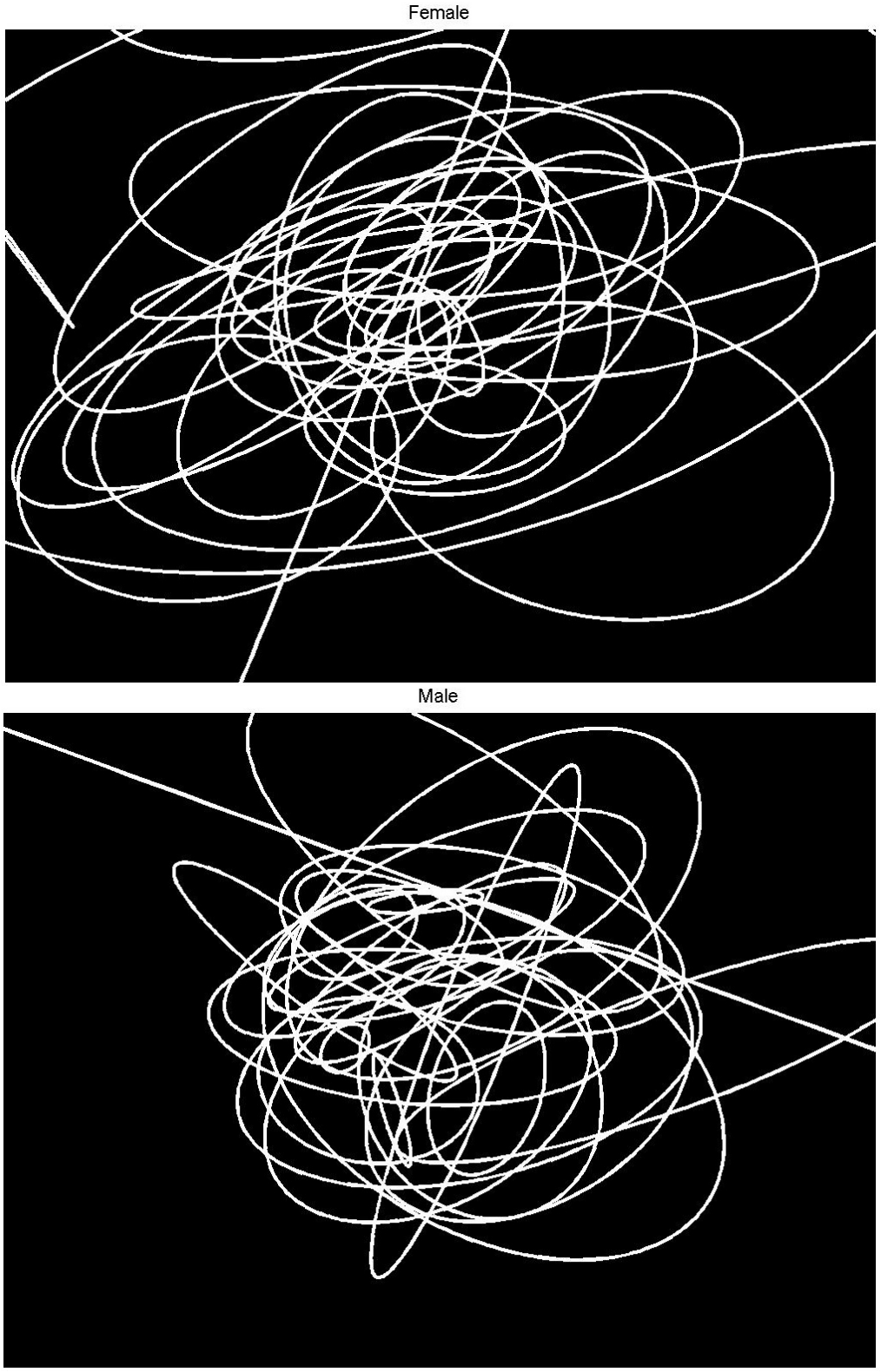

Figure 5.1: An example showing the covariance of the fixations between the male and female subjects. Center of ellipse is the mean, the shape and size is the covariance. The image (`r-025_0083.JPG`) shows a flowering cactus in the desert. The female subjects have more variance in the horizontal axis, σ_x.

due to a lack of prior literature on the eye-movements and religiosity, we are unable to provide probable explanations without further investigations with cognitive psychologists.

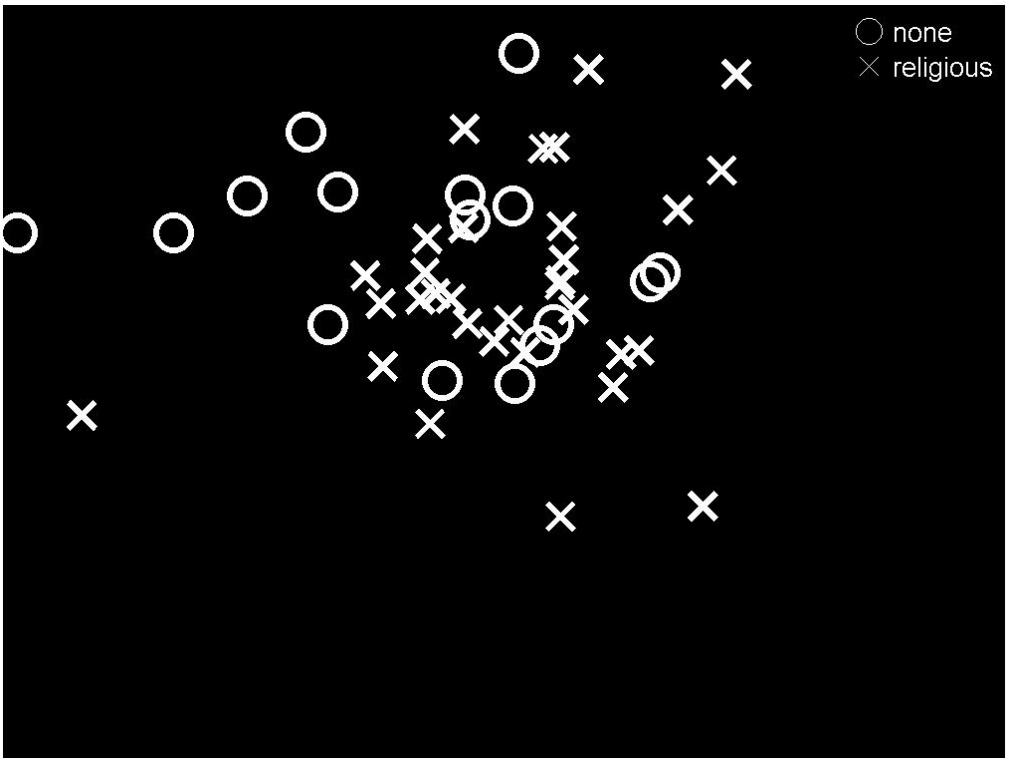

Figure 5.2: An example showing the mean of the fixations between the religious and non-religious (none) subjects. The image (`9606.JPG`) shows a simple wooden pail. The religious subjects fixations are more centrally aligned in the horizontal axis ($\bar{x}$) than non-religious subjects.

For *Extrovert/Introvert* groups, only the σ_{xy} is found to be significant. For *Sensing/Intuition*, the various fixation durations features, $\bar{d}$, σ_d, d_2, d_L, D, correlates positively with the *Sensing* group. This corresponds well with the characteristic of *Sensing* type who will spend more time to examine a stimuli before making a judgment.

In *Thinking/Feeling* groups, the $\hat{p}$ feature is a good indicator for emotions and it correlates positively with the *Feeling* group. In summary, the correlation analysis are reasonable and consistent with prior knowledge.

	Male/ Female	Religious/ None	Extrovert/ Introvert	Sensing/ Intuition	Thinking/ Feeling
$\bar{x}$	1	<u>17</u>	2	0	6
$\bar{y}$	1	2	0	7	6
d	0	5	0	<u>16</u>	3
σ_x	<u>18</u>	3	3	3	0
σ_y	<u>8</u>	6	6	2	2
σ_{xy}	6	3	<u>8</u>	4	0
σ_d	0	3	1	<u>10</u>	4
$\hat{p}$	<u>14</u>	<u>8</u>	3	7	<u>20</u>
x_1	2	<u>17</u>	1	0	3
y_1	2	1	3	<u>9</u>	7
d_1	0	3	3	7	5
x_2	5	<u>9</u>	1	3	7
y_2	4	5	5	<u>8</u>	<u>11</u>
d_2	4	<u>8</u>	2	<u>13</u>	<u>9</u>
x_L	2	<u>12</u>	3	4	<u>12</u>
y_L	3	4	1	3	<u>14</u>
d_L	0	5	1	<u>12</u>	5
D	3	<u>68</u>	0	<u>12</u>	6
N	3	<u>46</u>	2	4	7

Table 5.1: Correlation Analysis. The values in the table shows the number of images which $p - value < 0.05$ (*statistical significant*) for the feature. The features which have less than 7.5 ($0.05 * 150$) *statistical significant* images are considered to be statistical coincidences, and are not selected. The features which are selected as underlined.

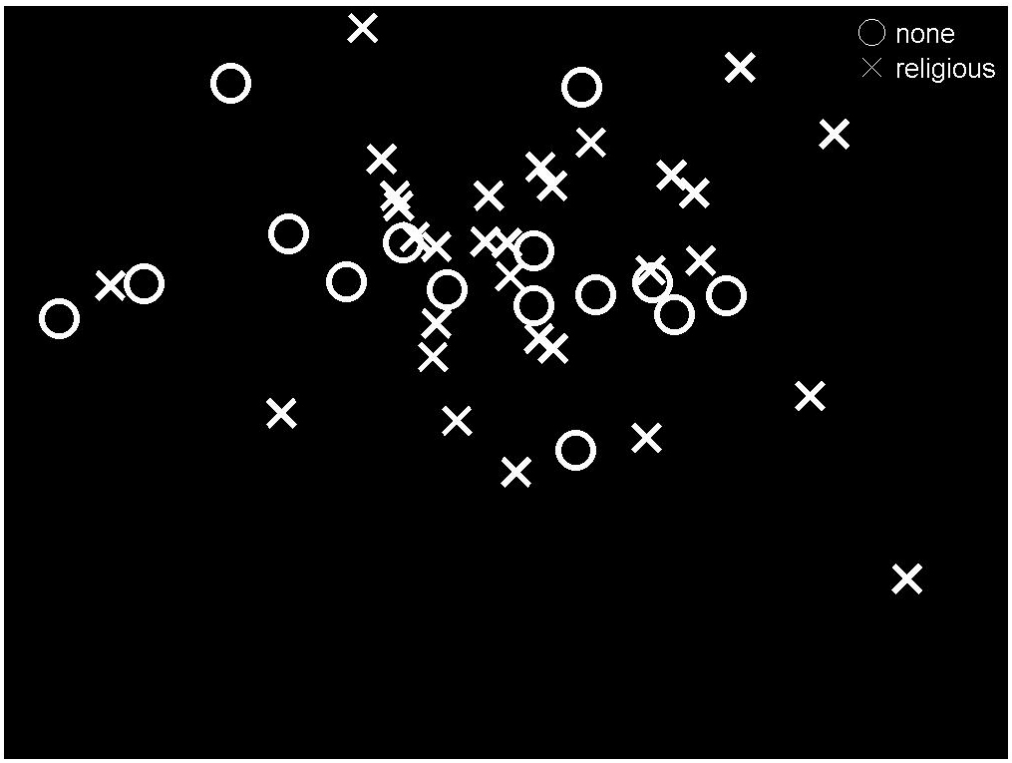

Figure 5.3: An example showing the first fixations between the religious and non-religious (none) subjects. The image (`9606.JPG`) shows a simple wooden pail. The religious subjects fixations are more centrally aligned in the horizontal axis (x_1 than non-religious subjects.

5.3 Classifier and training

We used the standard linear SVM classifier in the Matlab Biometric Toolbox, with the default parameters and auto-scaling.

For cross-validation, we applied the leave-one-out method. This method is most suitable as there are insufficient number of subjects per image for k-fold cross-validation, or train-validate-test division.

5.4 Empirical results and analysis

The Table 5.2 summarizes the experimental results. Accuracy is defined as the ratio of classification results matching the subject's answers. For example, the gender of the subjects are male, female, male, female and the classifications are male, male, male, female, then the accuracy is 0.75. This

	prior	mean	max	images
Male/Female	0.52	0.54	0.75	94
Religious/None	0.63	0.58	0.78	46
Extrovert/Introvert	0.52	0.51	0.66	80
Sensing/Intuition	0.62	0.52	0.76	26
Thinking/Feeling	0.63	0.54	0.73	24

Table 5.2: Accuracy of the classifiers. Prior probability refers to the prior proportion of the majority group. In our dataset, there are 27 females and 25 males subjects, thus prior probability for gender is $27/52 = 0.52$. Images refers to the number of images which classifiers' accuracies are higher than prior probability.

is because there are 3 out of 4 matches; the 2nd classification (male) does not match with the subject's answer (female).

Except for gender, the mean accuracies are lower than the prior distribution. This validates our claim that only certain images are useful for certain trait classification. The semantic meanings and affectiveness of the each image differs considerably in the dataset. This also indicates that differences of gaze information of gender are greater and can be more accurately differentiated.

The maximum accuracies are higher than the prior distributions for all factors, indicating that it is possible to classify the factors for using those images respectively. Male/Female, Religious/None and Extrovert/Introvert classifiers have many images which have higher accuracy than prior probability. Thus it is easier to select an appropriate image to suit the application requirements for these traits. The Religious/None trait has the highest maximum accuracy (0.78) while the Extrovert/Introvert the lowest (0.66). This suggests that religiosity has more influence on eye-gaze information compared to Extrovert/Introvert trait.

5.5 Classification using eye-gaze from multiple images

We further conducted a set of experiments in which trait are classified by gaze information of *multiple* images. There are many methods of combining the classifiers from single image classification. We experimented on the voting and tree ensemble methods.

The voting ensembles classifier is implemented as follows. For each subject, the classification results from the single image classifiers vote for the final class. For example, if a subject viewed 5 images and the respective classifiers' results are male, male, female, female, female; then the final classification result is female (3 votes vs 2 votes for male). For this method, the selection of the classifiers is critical to the accuracy rate. The selected single-image classifiers should be also independent for high accuracy. There are 3 selection methods. Using the single best classifier: *single*, using all classifiers: *all* and using the top k classifiers, k is the optimal number of classifiers: *greedy*.

We also use construct a binary decision tree method where each internal node is the results of single-image classifier. The construction method is the ClassificationTree class from Matlab's Statistical Toolbox (Coppersmith et al., 1999).

The experimental results are shown in Table 5.3. Clearly, using multiple images outperforms even the best single image classifier. The *all* ensembles have the worst accuracies. This is consistent with our observations that only some images are suitable for trait inference. The *tree* ensembles are generally good and only a few images are required. Thus it is suitable for applications which the users may not be willing to view too many images.

	single	all	greedy	tree
Male/Female	0.75	0.58	<u>0.87</u> (3)	0.85 (3)
Religious/None	0.78	0.65	<u>0.88</u> (8)	0.84 (2)
Extrovert/Introvert	0.66	0.53	0.80 (12)	<u>0.92</u> (2)
Sensing/Intuition	0.76	0.52	0.80 (3)	<u>0.92</u> (5)
Thinking/Feeling	0.73	0.54	<u>0.90</u> (13)	<u>0.90</u> (4)

Table 5.3: Mean accuracy of the multiple image classifiers. For *greedy* and *tree*, the number in the parentheses indicate the number of classifiers selected. The best accuracies for each factor are underlined.

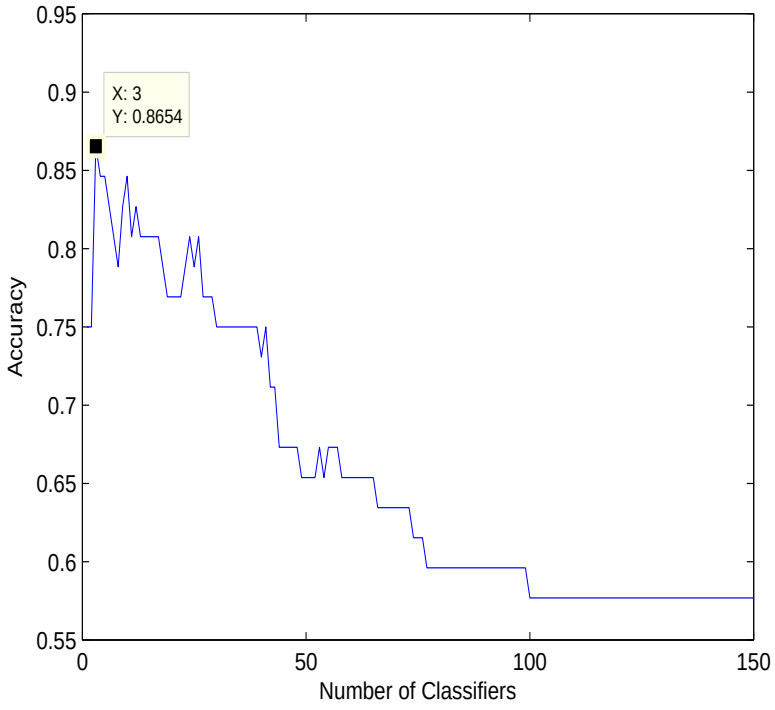

Figure 5.4: The accuracy plot for the greedy ensemble of gender classifiers. The box shows the optimal number of classifiers (3) which achieves the accuracy of 0.865.

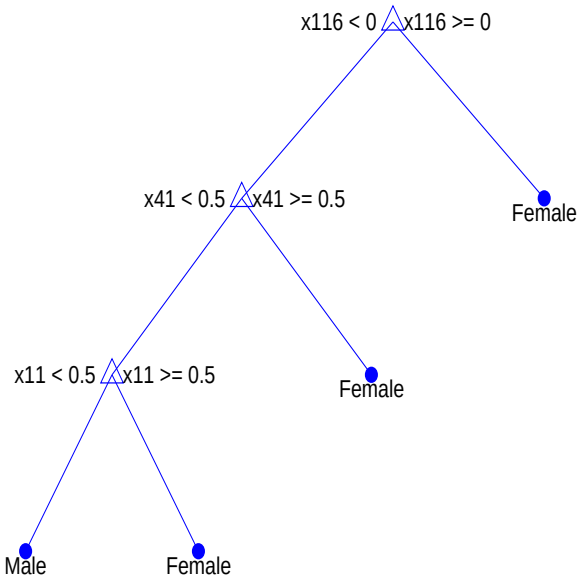

Figure 5.5: The decision tree ensemble of gender classifiers. xN refers to the single image classifier id and the single image classifier assigned -1 to Male and 1 to Female.

5.6 Discussions

The selection of the image or video which will elicit different eye-gaze patterns of the targeted group is an interesting multidisciplinary problem which will require knowledge of multimedia processing techniques, cognition and statistics. One possible future direction is to leverage on the widely reported Kosinski et al.'s research which links the "like" on Facebook with various demographic traits (Kosinski et al., 2013).

The interaction and influences of various traits on eye-gaze is another interesting and challenging problem to be solved. For example, both Female and Feeling subjects correlates positively to the $\hat{p}$ feature. Hence, potentially complicating the correct inference of a Male/Feeling subject.

The feature selection method and the classifier algorithm are the standard methods in machine learning. The purpose of this work is to establish that eye-gaze can be used for inference of P. As such, comparison of the dif-

ferent classifiers and their parameters on my applications while interesting, is outside the scope and is a potential future work.

The other eye-gaze features such as saccades and scanpaths may be useful features for solving this problem and would be investigated if resources are available.

Chapter 6

Trait-specific Fixations Prediction

`Prediction is very difficult, especially if it's about the future.`

`- Nils Bohr, Nobel laureate in Physics`

Fixation or saliency prediction is an important problem in saliency research. The objective is to develop a computational model to accurately predict human's fixations for a given visual stimulus, e.g. image. Applications are wide-ranging; from effective web-page design to features computation in computer vision and computer graphic problems. It is a particularly challenging problem as there are many factors which can affect eye-gaze. Current research focus on V (e.g. edges, colors), I (e.g. tasks) or VI models (See Chapter 3 for more details).

To the best of our knowledge, the P factors are not incorporated in any existing fixation predictors (Frintrop et al., 2010). With our VIP model, we can clearly identify this gap and propose the VP model for predicting fixations. In our VP model, we assume that the I factors are constant for all subjects. For our experiment, I is the "free-viewing" task.

In the bottom-up fixation prediction problems, the objective is to develop an algorithm to accurately predict the new fixations given a stimulus. The human eye-fixations are considered the ground-truth. In other words, a predictor trained with fixations from other viewers is the benchmark which other proposed algorithms are compared against. Hence, for any given stimulus, if VP predictor outperforms V predictor, then any existing bottom-up fixation prediction algorithm (V model) can expect to have better accuracy by incorporating the relevant P factor(s).

Claim: If a trait has an influence on fixations for a given image, then it is expected that a VP predictor will have higher accuracy compared to one which is trait-agnostic (V predictor), *all other conditions being equal.*

Using gender as an example trait, the V model assumes that for *any* stimulus, a predictor trained with only male subjects' fixation data will be equally accurate for predicting males' and females' fixations. The VP model assumes that for *some* stimuli, a predictor trained with only male subjects' fixation data will be more accurate for predicting males' fixations than females'. From prior work, some categories of gender-divergent stimuli would be faces (Ramanathan et al., 2010), nude images (Ramanathan et al., 2010), conversation video (Shen and Itti, 2012) and movies (Goldstein et al., 2007).

6.1 VIP formulations

There are 2 types of fixation prediction problems for images. The first type of problem is to predict fixations for any image, while the second type is to predict fixations for the images with training data. For our experiment, we consider the second problem type. That is given an image and its training data, we would like to predict as accurately as possible the fixations of an unknown subject on the image.

A VP predictor is trained from eye-gaze information from subjects with the same trait (e.g. $gender = female$). For a test subject with trait = $p1$, the $p1$-specific predictor will be applied. Similarly, a test subject with trait = $p2$, the $p2$-predictor will be applied.

The VP predictor is defined as: $\underset{I=c}{h}(E_p)$ where $\{E_p\}$ is the set of E from training subjects with $P = p$, e.g. $gender = female$.

As an example, if the test subject is a female, the female-specific predictor which is trained with fixations from female subjects, will predict the fixations.

The V predictor does not include the P factors and each subject is indistinguishable from another. This predictor defined as $\underset{I=c}{h'}(\{E\})$ where $\{E\}$ is the set of the fixations for all available training subjects. In other words, V predictors are agnostic to traits.

6.2 Experimental setup

We select the Gaussian filter for $\underset{I=c}{h'}(\{E\})$. It is the standard predictor function for fixation prediction as it has the desirable property of smoothing inherent spikiness of the fixation data. The size and sigma parameters are dependent on the mapping of the pixel to the degree of vision (Judd et al., 2012). For our experimental setup (See Chapter 4.1.1), we used 80 pixels for size and 20 for *sigma*. These values approximates to 2° and 0.5° respectively. Human's vision has highest acuity within 2° and fall offs rapidly from there.

The output of the Gaussian filter is a saliency map from which the fixations of the test subjects are compared against using the area-under-the-ROC-curve (AUC) measure. The AUC is a standard measure in fixation prediction research (Borji and Itti, 2013). First, the ROC curve is obtained by plotting the test subject's fixations on the saliency map. The non-zero

Traits	Majority	Minority
gender	Female (0.52)	Male (0.48)
agegroup	24 and below (0.71)	25 and above (0.29)
ethnicity	chinese (0.71)	non-chinese (0.29)
religion	religious (0.63)	none (0.37)
specialty	science and engineering (0.71)	others (0.29)
education	tertiary (0.77)	post-graduate (0.23)
income	0-999 (0.77)	1000 and above (0.23)
expenditure	1-999 (0.77)	1000 and above (0.23)
nationality	non-singapore (0.52)	singapore (0.48)
ei	Intovert (0.52)	Extrovert (0.48)
sn	Sensing (0.62)	Intuition (0.38)
tf	Feeling (0.63)	Thinking (0.37)

Table 6.1: Grouping of traits for the VIP dataset. The numbers in parentheses show the distribution of the traits. Nationality and country has same distributions, and are combined into 1 trait.

values on the saliency map on which the fixations occur are "true positive" and the rest of the non-zeros in the saliency map are "false positives". The ROC curve plots the "true positives" (sensitivity) vs. "false positives" (1 - specificity), for a binary classifier system as its discrimination threshold is varied. A predictor which produces random results will have an AUC of 0.5 and an ideal predictor will have a AUC of 1.0.

We used the VIP dataset for our experiments, the task is $c =$ "$free - viewing$". Similar to experiment for personal trait inference (Chapter 5), we pre-select the 52 subjects with good quality of eye-gaze data. For fair comparison, we did not pre-process the fixations to remove outliers. We noted that there are some images for which no fixation data is recorded for some subjects within the region of the presented stimulus. This could be due to eye-tracking errors or the subjects being distracted during the presentation of the stimulus.

We group the traits into 2 classes for a more even distributions such that there are sufficient training samples for the traits. Table 6.1 shows the grouping.

For validation, we use the leave-one-subject-out method. With this

method, for each image, a subject's fixations are predicted from the saliency map computed from other *selected* training subjects' fixations. The AUC is computed from the subject's fixations and from the saliency map as described above. This is repeated for every subject. An average AUC score is then obtained by averaging the AUC scores for all subjects. We refer this average AUC score for all subjects as $AAUC$ for the rest of this chapter. For every image, there is an $AAUC$ for VP predictor and an $AAUC$ for V predictor.

For the VP predictor, for each image, the saliency map is generated from the training subjects with the *same* trait as the test subject. For the V model, for each image, the saliency map is generated from the training subjects with the *different* trait as the test subject. If trait is a factor in fixation prediction for an image, then we would expect the $AAUC$ for VP predictor be higher than V predictor's $AAUC$. We used subjects with *different* trait for V predictors instead of using all subjects for training so that the number of training samples are similar for both predictors. The impact of training sample size is discussed in Chapter 6.4.

6.3 Empirical results

We compute the mean $AAUC$ on *all* images for using VP predictors as shown in Table 6.2. For *gender*, *religion* and the 3 personality types, the V predictors have slightly higher mean $AAUC$. The maximum difference is only 0.003 and the mean difference is 0.0017. For the remaining 7 traits, the VP predictors have significantly higher mean $AAUC$. The maximum difference is 0.03 and the mean difference is 0.017. Therefore, VP is better for than V for more traits (7 vs 5) and the differences in mean $AAUC$ is 10 times higher.

More importantly, the advantage of our VP approach is shown in Ta-

Trait	$\overline{AAUC}_{VP}$	$\overline{AAUC}_{V}$	$\overline{AAUC}_{VP} - \overline{AAUC}_{V}$
gender	0.752	0.754	-0.002
agegroup	0.753	0.739	0.014
ethnicity	0.755	0.748	0.006
religion	0.752	0.752	-0.000
specialty	0.754	0.727	0.027
education	0.757	0.733	0.024
income	0.754	0.736	0.018
expenditure	0.754	0.724	0.030
nationality	0.753	0.751	0.002
ei	0.753	0.755	-0.003
sn	0.751	0.754	-0.003
tf	0.752	0.753	-0.001

Table 6.2: Comparison of the mean $AAUC$ for the VP and V predictors for all images.

Trait	$\sum AAUC_{VP} - \sum AAUC_{P}$
gender	-0.34
agegroup	2.11
ethnicity	0.95
religion	-0.03
specialty	4.06
education	3.60
income	2.68
expenditure	4.53
nationality	0.30
ei	-0.44
sn	-0.42
tf	-0.08

Table 6.3: The table shows the net differences in $AAUC$ between VP and P predictors.

Trait	$\|AUC_{VP} > AUC_V\|$	$\|AUC_V > AUC_VP\|$
gender	62(0.41)	88(0.59)
agegroup	113(0.75)	37(0.25)
ethnicity	89(0.59)	61(0.41)
religion	72(0.48)	78(0.52)
specialty	130(0.87)	20(0.13)
education	128(0.85)	22(0.15)
income	120(0.80)	30(0.20)
expenditure	132(0.88)	18(0.12)
nationality	81(0.54)	69(0.46)
ei	61(0.41)	89(0.59)
sn	56(0.37)	94(0.63)
tf	70(0.47)	80(0.53)

Table 6.4: The table shows the number of images which VP predictors which have higher AUC than the V predictors, and vice versus. The numbers in parentheses are the ratio. The total number of images is 150.

ble 6.3. There are significant net gains are $AAUC$ by using VP predictors for 7 traits while the net losses are comparably lower for the 5 traits. In other words, when VP predictors are better, they are much better; and when they are worse, they are just slightly worse than V predictors.

6.3.1 V factor

From our experimental results, we ascertain that some selected images can elicit trait-divergent fixations. That is these images elicit different fixation patterns for people with different traits. Table 6.4 shows the number of images which are trait-divergent. The portion of images which are divergent are quite high for the VIP images; ranging from 0.37 (*ei*) to 0.88 (*expenditure*). If we can extrapolate this observation to other visual stimuli, then the probabilities that any image is trait-divergent is quite high for many traits.

For example, although *gender* VP predictors are worse than V predictors on average, we can easily find as many as 62 out of 150 images which the VP predictors are more accurate than the P predictors. One example shown in Figure 6.1 and Figure 6.2. This image elicits different fixation

patterns from male and female subjects. The VP predictors' *AAUC* is 0.024 higher than the V predictors' *AAUC*.

It is certainly of research interests to systematically identify the properties of images which are contributing to the differences in fixations. However, this intrudes significantly into psychology research and is beyond the scope of this thesis. Nevertheless, from prior psychology research, we know that faces are some of the most important objects for many traits (Shen and Itti, 2012; Wu et al., 2013).

6.3.2 P factor

From Table 6.4, We noted that the 4 traits: *specialty*, *education*, *income* and *expenditure* have at least 80% of images with which VP predictors are more accurate. These 4 traits are all proxy metrics for social-economic status (SES). SES is an economic and sociological combined total measure of a person's work experience and of an individual's or family's economic and social position in relation to others, based on income, education, and occupation. Our experimental results indicate that this factor modulates fixations on more images in our dataset than other more well-studied traits, such as gender (Goldstein et al., 2007; Shen and Itti, 2012), age (Goldstein et al., 2007) and personality (Risko et al., 2011; Wu et al., 2013).

There is a vast wealth of literature on the effects of SES on health, language development, parental interactions and non-verbal interactions. Interestingly, to the best of our knowledge, there is very little research on the relation between eye-gaze and SES. The only reference we found is a study by Hunziker on intelligence and eye-gaze (Hunziker, 1970). On one hand, our experiment may inspire other researchers, especially visual perception psychologists, to conduct more studies on this interesting observation. On the other hand, we are unable to strongly support our observation of the modulating effects of SES on fixations on many images with prior studies.

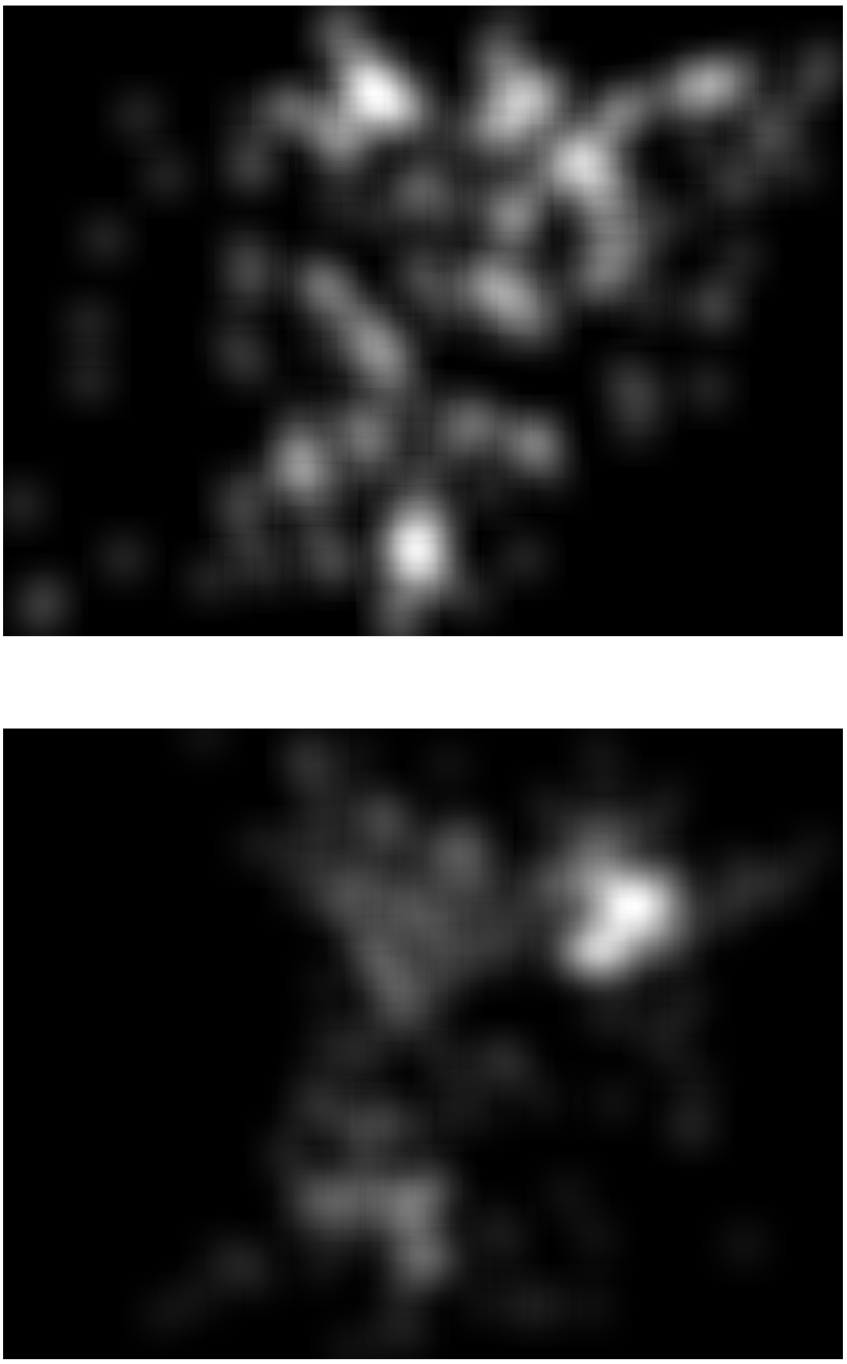

Figure 6.1: Gender-specific saliency maps. The maps are generated by applying the Gaussian filter on fixations of every subject from the respective genders for the 7192.*jpg* image. The top is the female saliency map and the bottom is the male saliency map. The female subjects fixated more on the bottom and are more spread-out while the male subjects more on the top-right region.

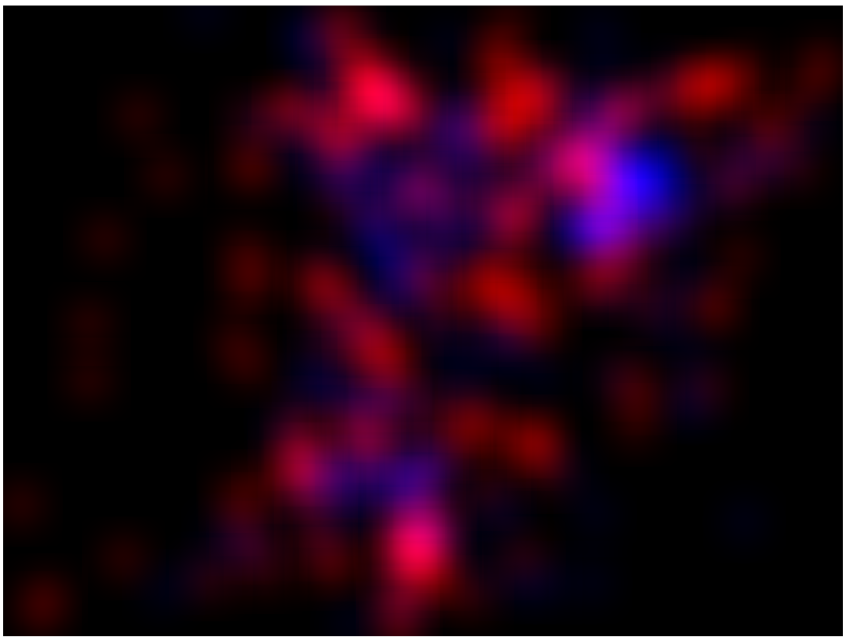

Figure 6.2: This figure is best viewed in color. Gender-specific saliency maps. The maps are generated by applying the Gaussian filter on fixations of every subject from the respective genders for the 7192.*jpg* image. The red channel is the female saliency map and the blue channel is the male saliency map. The purple regions are where the genders are both fixated on.

6.4 Discussions

Our experimental results show that given the choice between collecting trait-specific data and trait-agnostic data, using trait-specific data will produce significantly higher accuracy for many traits. Furthermore, for traits such as gender, religiosity and personality types, the V predictors are only slightly more accurate on average. These 2 observations made VP predictors a better choice for many applications.

We consider the qualifying statement of our claim: *all other conditions being equal*. One important condition is the amount of training data. Given an existing set of training data, is it better to use the whole data set for training the predictor, or to use only the subset of trait-specific training data?

We implemented a predictor, U, which is trained with all subjects' fixations, except the test subject's. In effect, this U predictor has double the amount of training data over the VP and P predictors. Using *gender* as an example, the female-specific predictors will have 26 training samples, and male-specific predictors will have 24 training samples. In comparison, the U predictor has 51 training samples.

The mean $AAUC$ for the U predictor is 0.773 which is higher than the V or VP predictors' mean $AAUC$ (Table 6.2). This indicates that the number of training examples have greater impact on the accuracy than using a trait-specific approach. In general, researchers should use all available training data rather than using only a subset of trait-specific data. However, for some specific stimuli, using all training data may actually be slightly worse. For example, there are 3 images which have slightly higher $AAUC$ with our VP predictors than the U predictor.

To summarize the above, it is better to use as much training data as possible. However, when acquiring training data, it is more efficient to acquire data which are traits-specific for the application. For example, to

predict fixations for a website which targeted audience are women, conscious efforts should be made to collect training data from female adults. However, if there is already existing fixation data from the general population, we should use the whole set rather than selecting only the female adults' data.

Our experimental results support our claim that P factors which are currently ignored by saliency researchers contribute to the differences in fixations for *some* stimulus. Indeed, our experimental results are consistent with recent psychological studies (Shen and Itti, 2012; Risko et al., 2011).

We hope that these experimental results will inspire other researchers to re-examine their assumptions about the importance of the P factors with regards to human visual attention and eye-gaze research. We strongly believe that the current computational models are incomplete without considering the various P factors. An intuitive example would that the saliency map of a nude image for male (predominately in the sexual organs regions) and female observers (predominately in the facial regions) will be very different and therefore should be modeled differently.

Chapter 7

Implicit just-in-time profiling

`The Eyes are the window to your soul`

`- William Shakespeare (Renowned Playwright)`

In this chapter, we will present an implicit just-in-time profiling method from eye-gaze when viewing video. The accuracies of the inference are greatly improved by exploiting the temporal dependency on the shots in the videos. `Eye-2-I` is also able to infer emotions and topics of interest, besides personal traits. This makes `Eye-2-I` the first complete VIP system.

7.1 Sample scenario

Alice, our hypothetical user, likes personalized services and contents, as long as her privacy is not compromised (survey by Adobe Systems and Edelman Berland (2012)). She understands that for personalization to work well, profiling of her demographic, socioeconomic, psychographic, geographic, lifestyle and interests information is necessary (Google, 2014b; Yankelovich and Meer, 2006; Qiu et al., 2012). Existing methods suffer from **data impurity**, **privacy and security issues** and **need for historical data**.

Data purity. Alice shares her tablet device with her family; and she works on her personal laptop computer. Because of these reasons, Google's tracking profiling method from her online behavior, such as browsing, search and account activities over many days is not accurate (Google, 2014b). She is harassed by barrage of men's deodorant ads on her tablet (contaminated profile resulting from her brother's browsing history); and eye-tracking ads on her personal laptop (out-of-context profiling from her work search history for eye-tracking topics).

Privacy and security. Facebook, using questionnaires and account activities to profile Alice, did not fare much better (Facebook, 2014). She is among the one-third of social-networkers who provided false information as she is very concerned with her privacy (emedia, 2007). After her search history and account information from other companies were leaked (Kawamoto and Mills, 2006) and stolen(Gibbs, 2013), she lost her trust in online storage of her personal data.

Historical data. Alice flies frequently and she finds that the content provided by the airports' public interactive kiosks is never personalized. Since these systems do not have her history, personalization is not currently possible.

Alice is clearly not satisfied with the situation. Current profiling methods using tracking data or questionnaire require historical data to be acquired and stored. In the acquisition processes, personal data from multiple users may be mixed together; data from different context may be mixed; or user may simply provide incorrect data. Secure and privacy preserving storage of personal data is challenging and expensive. These problems can be avoided if implicit profiling is done on-the-fly and be usable within minutes, i.e., just-in-time.

Hence, we propose a novel implicit just-in-time profiling method will help these businesses to serve Alice better.

7.2 Proposed method

Our inspiration is from implicit tagging where the meta-data about a multimedia content is derived from the observer's natural response (Soleymani et al., 2012). Instead of tagging the multimedia content, our method infers the observer's emotions, traits and interests from the natural responses: eye-gaze and optionally face. In other words, our proposed method "tags" the users from their interaction with the system.

Eye-gaze is an automatic and subconscious behavior. It is well-established that eye-gaze is influenced by a person's interests (Barber and Legge, 1976); and pupil dilations correlate with emotional arousal (Bradley et al., 2008). Recent psychology research finds that eye-gaze is also modulated by personal traits. Chua et al. measured the gaze differences between American and Chinese participants. Americans fixated more on focal objects on a complex background and they tend to look at the focal object more quickly (Chua et al., 2005). Goldstein et al. examined the viewing patterns when watching a movie and observed that male and older subjects were more likely to look at the same place than female and younger subjects (Goldstein et al., 2007). Wu et al. discovered that the personality relates to fixations towards eye region (Wu et al., 2013).

To the best of our knowledge, we are the first to propose inferring a comprehensive profile of emotions, demographic traits, personality types and topics of interest from eye-gaze, and optionally face.

Data purity. As the profiling is performed just-in-timely, there is no risk of contamination from prior history. We can therefore assume that the eye-tracking data is valid, relevant and up-to-date. While eye-tracking research is still progressing at a steady state, we regard current state-of-the-art algorithms are sufficiently accurate and precise for profiling as demonstrated in our experiments.

Privacy and security. The profiling can be performed either remotely

or locally. The profile can either be reused or disposed off immediately after the session ends (transient). If the profiling is local and transient then the risks for both issues are greatly mitigated. Our method has the flexibility of being either a secure and privacy preserving method; or being an `just-in-time` and `implicit` method which complements current profiling methods.

Historical data. By eliminating the requirement for user's historical data, our profiling method can be used in new applications which are impossible with current methods, e.g. public interactive kiosk. Our method assumes the availability of eye-tracking data. This data can be captured from a standard video camera. As most modern devices, including laptops, tablets, smart-phones, gaming consoles and smart televisions are equipped with a front-facing video camera, it is a realistic assumption. A specialized eye-tracker, if available, will provide higher eye data fidelity. As eye-tracking technology matures, we expect the availability of eye-trackers to increase significantly in the near future (Rosenberg, 2006).

7.3 Related work

As users browse Google's partner websites, it stores an advertising cookie in a user's browser to understand the types of pages that user is visiting. This information is used to show ads that might appeal to the users based on their inferred interest and demographic categories. Google may also use demographics from Google users' profiles (Google, 2014b).

Questionnaires is a simple and direct method for profiling. It suffers from response time, obtrusiveness and inaccurate entries (intentional or otherwise). In comparison, our method is just-in-time, implicit and behavioral-based and hence more suitable for applications such as targeted advertising.

Attributes	Eye-gaze	Face	Our method
emotion type		Gevers (2014); Gunes and Piccardi (2009)	$F+V$
emotion intensity	Barber and Legge (1976)	Koelstra and Patras (2013)	$E+F+V$
gender	Goldstein et al. (2007); Ma et al. (2013)	Buchala et al. (2005)	E
age	Goldstein et al. (2007)	Buchala et al. (2005)	E
ethnicity/culture	Chua et al. (2005)	Buchala et al. (2005)	E
personality	Wu et al. (2013); Ma et al. (2013)	Martens (2012)	E
religiosity	Ma et al. (2013)		E
interests	Barber and Legge (1976)		E
field of work/study			E
education			E
socioeconomic			E

Table 7.1: Comparison of the attributes which are correlated and/or inferred with eye-gaze, face and our proposed system: `Eye-2-I`. F are face features, E are eye-gaze features and V are visual-audio features.

Facial features provides an alternative mean of profiling implicitly and just-in-time. Personal traits such as gender, age and ethnicity can be inferred from facial features (Buchala et al., 2005). However, our method can also be used to predict other demographic factors which may not manifest in appearance-based methods, e.g. religiosity. Another clear advantage of using eye-gaze is that transient mental states such as topics of interest can be revealed through interactions between the eye-gaze and specific regions in the video content. Table 7.1 shows the comparisons between the 2 modalities. More importantly, the two methods are complementary and are included in our proposed system.

An important research is on improving the sub-category of contextual targeting. Contextual targeting is related to the content and does not use user's prior data. In VideoSense (Mei et al., 2007) and AdImage (Liao et al., 2008), contextual relevance is computed from textual meta-data and video analysis algorithms. CAVVA computes the affective of the video from visual and audio analysis (Yadati et al., 2014). Besides providing contextual information, these algorithms also identify the optimal points in a video to insert the video-in-video advertisements. For example, CAVVA employed a non-linear optimization function and a genetic algorithm based solution to identify the advertisement insertion points and select the correspond-

ing advertisements in a unified manner. Our proposed system, `Eye-2-I` contains a sub-module which analyzes the affect in videos. It also provides other contextual data from the content provider and from an optional user's feedback process.

7.4 System design

To demonstrate our proposed method, we designed the eye-gaze profiling for targeted advertising in videos, `Eye-2-I`. This application has some attractive properties for a proof-of-concept system. In targeted advertising systems, inaccurate profiles will not cause catastrophic outcome so extremely high accuracy is not mandatory. Watching videos is a very popular activity and is well-accepted by users for targeted advertising, e.g. YouTube (Adobe Systems and Edelman Berland, 2012). Furthermore, the natural and implicit interactions between the video content and the users generate much eye-gaze data for the profiling system. The temporal ordering of the shots within a video enable us to employ incremental classification method which substantially improves the accuracy over per-shot classifications. Video watching is a self-containing activity and thus suitable for local transient profiling; thereby being secure and privacy preserving.

A typical video targeted advertising system consists of 2 sub-systems: user profiling and advertisement selection. `Eye-2-I` infers the emotions, traits and interests of the viewer. A suitable selection process then matches this information against many external factors including purchased keywords, negative keywords, target market segments, product categories, advertisements' content, delivery schedules, bandwidth limits, click-through rates and bidding process etc to select the set of advertisements to be presented (Rajaraman and Ullman, 2012). Figure 7.1 shows the system diagram of `Eye-2-I`.

	Eye-2-I			Ads Selection
	Input	Module	Output	
video	content's provider	Meta-data	descriptions	Contextual
			keywords	
			genre	
	[feedback]*		comments	
			ratings	
shot	shot	Affect Analysis	valence	
			arousal	
	[face]*		expressions	
	eye-gaze*		dilations	
		Interests Inference	topics of interests	Interests-based
		Traits Profiling	demography	Placement
			personality	

Table 7.2: Mapping of input, module and output of `Eye-2-I`. The first column indicates the scope of inputs/outputs, i.e. per video or per shot. * denotes modalities from the users. The inputs in brackets are optional.

Since our objective is to demonstrate the applicability of our profiling method as inputs to targeted advertising, the advertisement selection sub-system is beyond the scope of this thesis. Our experiments are designed for the profiling sub-system only.

Advertisement targeting system can be sub-categorized into contextual targeting, interest-based advertising and placement targeting (Google, 2014a). Contextual targeting displays advertisements related to the content. Interests-based advertising are based on users' interests. Placement targeting are based on demographic and geographical locations etc. `Eye-2-I` extracts and returns meta-data and performs affect analysis for contextual targeting; infers topics of interest from eye-gaze for interests-based advertising; and infers demographic and personality traits for placement targeting. Table 7.2 shows the mapping of `Eye-2-I`'s outputs to these sub-categories.

In designing the `Eye-2-I` system, we carefully consider the security and privacy issues. As such the system is self-containing and can be implemented on an offline device. The inference of user's emotions, traits and

interests are only based on a single video, there is no use for any prior data of the user. For added security, the inferred data can be discarded as soon as a video ends or when a user's session ends.

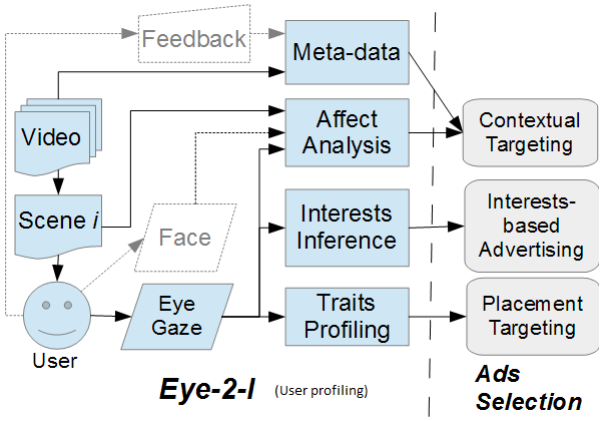

Figure 7.1: The system diagram for the `Eye-2-I`. The dimmed modules are optional. The video is segmented into multiple shots, each consisting of a series of uninterrupted frames. For each shot, the user's facial expressions (optional) and eye-gaze data are analyzed for affects, interests and personal traits. Valency and arousal scores are pre-computed for each shot. Textual description, keywords and genres of the video, provided by the content's provider, is extracted from the video. Optionally, meta-data from anonymous user's feedback on the video such as ratings and comments can be made available to an Ads Selection system. The affect analysis module processes data from 3 multi-modal sources: video content, eye-tracking data and video camera.

A typical session with the `Eye-2-I` starts with the sending of meta-data of the selected video to the Ads Selection system. While the user is viewing the video, the system tracks the eye-gaze and analyzes the facial expressions. At the end of each shot, which lasts for an average of few seconds, the system returns the inferred profile (emotions, demography, personality and interests) to the Ads selection system. According to the profile and the meta-data, the selection system will decide to either present a set of personalized advertisements or let the user continue watching the video without interruptions. The decision making process is similar to any existing profiling system which used historical data of the user, with the additional benefits that the inferred profile is up-to-date and not contaminated by other users or context. At the end of the viewing, the system will

demography	personality	interests
(0.87) male	(0.83) introvert	(0.54) computer
(0.70) > 25 y.o.	(0.90) sensing	(0.53) games
(0.66) Chinese	(0.81) thinking	(0.82) travel
(0.76) no religion		
(0.68) sci&eng	**affect analysis**	
(0.79) postgrad	(0.7) happy	(0.3) surprise
(0.57) < $1000	A = 73	V = 56
(0.53) < $5000	pupil: $226.45mm^2$	

Table 7.3: Sample output of `Eye-2-I`. The number in the parentheses indicates the confidence level, computed from the training data.

discard the profile unless it is needed for other purposes. The user may be optionally prompted to provide feedback to the system. Table 7.3 shows a sample output of `Eye-2-I`

7.5 Problem formulation

Our *VIP* framework characterizes computational eye-gaze research. It states that eye-gaze is a function of Visual stimulus, Intents and Personal traits. This maps naturally to the 3 subcategories of targeted advertising. Clearly, contextual targeting is dependent on Visual stimulus; interests-based targeting on Intents; and placement targeting on Personal traits.

We formulate our application as:

$$\{I, P\} = f^{-1}(E, V) \tag{7.1}$$

That is for a video shot (V), our `Eye-2-I` algorithms: f^{-1} infers interests, emotions (I) and personal traits (P) from the eye-gaze data (E) and offline content analysis.

We claim that we are the first to implement a complete *VIP* system.

7.6 Traits and interests profiling

From many psychological studies, it is clear that eye-gaze is correlated with various attributes: demographic profile, personality and interests (Goldstein et al., 2007; Chua et al., 2005; Wu et al., 2013; Barber and Legge, 1976). This suggests that automatic classification of these attributes from eye-gaze is not only feasible but also scientifically grounded. In these studies, one important factor is the visual stimulus. The type of attributes which are correlated with eye-gaze is `strongly` dependent on the properties of the visual stimulus. This in turn implies that the attributes which can be inferred from the eye-gaze is shot-dependent. Therefore, it is unlikely that any single shot will be useful for inferring all attributes from eye-gaze; and conversely, an attribute can be accurately inferred independently of the visual stimulus. A video shot refers to series of frames which runs for an uninterrupted period of time.

We identify the following personal traits for profiling: gender, age-group, ethnicity, religion, field of study/work, highest education qualifications and income groups (personal and household). Many of these traits are used in market segmentation and targeted advertising (Google, 2014a).

Jacob et al. found that advertisements were evaluated more positively the more they cohered with participants' personality types (Hirsh et al., 2012). `Eye-2-I` infers the Carl Jung's personality types: Extrovert/Introvert, Sensing/Intuition and Thinking/Feeling for the eye-gaze (Jung et al., 1991).

For the inference of interests, we have selected the same set of categories as Google Ads system, shown in Table 4.6. These categories are generic and does not match precisely to any specific product, e.g. Nike Shoes. Instead of using a vision-based approach, e.g. object detection, the inference is performed with only eye-gaze data.

7.6.1 Features extraction

The fixations are extracted from the eye-movements. From the fixations, 2 types of features are computed: statistical features and our novel region-of-interests (ROI) features.

Statistical features

The 19 statistical features are identified in Chapter 5 for inferring personal traits from eye-gaze viewing images (Ma et al., 2013). These features are found to be different among people with different traits from prior research. σ_x, σ_y and σ_{xy} for gender (Goldstein et al., 2007); x_1, y_1, d_1, σ_d, x_L, y_L, d_L for culture/ethnicity (Chua et al., 2005); $\hat{p}$ for personality type (Thinking/Feeling) (Bradley et al., 2008). We are the first to apply these features to infer personal traits and interests in video watching activity.

ROI features

Additionally, we propose the novel ROI features based on the observation that people with different interests will fixate in different ROI in a shot as reported by Barber and Legge (Barber and Legge, 1976). Our feature extraction algorithm quantifies the amount of attention given by a user in the different ROI of a given shot.

The algorithm first generates a saliency map by applying a Gaussian filter ($gaussianFilter$) to the training set of fixations . The size and sigma parameters are dependent on the mapping of the pixel to the degree of vision (Judd et al., 2012). For our experimental setup (See Chapter 4.1.1), we used 80 pixels for size and 20 for *sigma*. These values approximates to 2° and 0.5° respectively. Human's vision has highest acuity within 2° and fall offs rapidly from there. This is known as the foveal vision. The resulting saliency map models the amount of details perceived by the user for the shot. Alternatively, the saliency map can be automatically generated from

the shot using a suitable saliency prediction algorithm, e.g. our proposed VP predictors (Chapter 6).

The saliency map is then smoothed with a circular averaging filter (*pillbox*) of size 80 pixels to smooth out peaks which are too near to each other. The smoothed map is then normalized (*normalized*) between 0.0 and 1.0. Then the algorithm extracts the local maximum points ($localMaximumPoints$) from the normalized smoothed saliency map. These points are the centroids of the ROI in each shot. The pseudo-code for computing these points is shown in Algorithm 1.

The set of weighted distances are then computed between these points and a set of input fixations from a single user viewing the shot. The weighting is computed to simulate the foveal acuity of human eye. Visual information captures by the eye can be approximated as inverse exponent function of the euclidean distance ($euclideanDistance$) from the centroid of ROI (Hunziker, 2006). These weighted distances are multiplied by the respective fixation's duration to obtained the quantified attentional value given by each fixation against the set of ROI. Finally, attentional values are summed over all fixations for the combined attention of a viewer to each ROI. The feature's length is equal to the number of ROI centroids, and each value represents the amount of attention paid to the respective ROI. Algorithm 2 shows the pseudo-code for computing the ROI feature vector from a set of input fixations and the precomputed ROI centroids.

The ROI feature vector concisely represents the amount of attention given to each ROI for a given shot. As a region-based feature, it captures the location of eye-gaze at a higher granularity than the statistical features. See Figure 7.2 for an example of ROI.

Algorithm 1 Compute locally most salient points from a set of fixations

Parameter: F_s: set of training fixations for a shot, s
Parameter: $HSize$: size of the Gaussian and pillbox filter, default to 80 pixels
Parameter: $Sigma$: sigma value of Gaussian lowpass filter, default to 20

function ROICENTROID($F_s, HSize, Sigma$)
  $SaliencyMap \leftarrow$ gaussianFilter(F_s,$HSize$,$Sigma$)
  $SmoothMap \leftarrow$ pillbox($SaliencyMap$,$HSize$)
  $NormSmoothMap \leftarrow$ normalized($SmoothMap$)
  $RPoints \leftarrow$ localMaximumPoints($NormSmoothMap$)
  return $RPoints$
end function

Algorithm 2 Compute the ROI feature vector

Parameter: $RPoints = \{c_j\}$: Centroid for regions of interests
Parameter: $F_s^u = \{f_i\}$: set of fixations of user, u, in shot, s

function REGIONFEATURES($RPoints, F_s^u$)
  for all f_i in F_s^u **do**
    for all c_j in $RPoints$ **do**
      $PairDist_{ij} \leftarrow$ euclideanDistance(c_j,f_i)
      $AttentionVal_{ij} \leftarrow$ duration(f_i)$*e^{-PairDist_i}$
    end for
  end for
  for all $AttentionVal_{ij}$ **do**
    $ROIFeatures_j \leftarrow \sum_i AttentionVal_{ij}$
  end for
  return $ROIFeatures$
end function

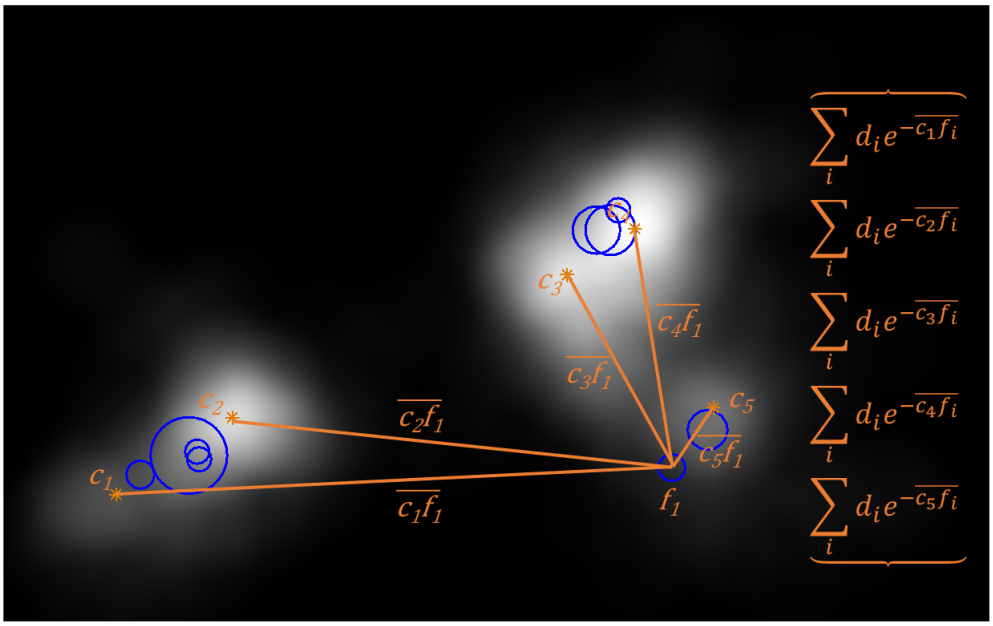

Figure 7.2: This figure is best viewed in color. An example of the ROI features. The background image is the saliency map computed from the fixations of all training subjects. The $*$ denotes the centroids, c_j, of each ROI region. The blue circles are the input fixations. For each fixation, f_j, the Euclidean distance, $\overline{c_j f_i}$, between the fixation and the centroid is computed. The inverse exponential function, $e^{-\overline{c_j f_i}}$ is weighted by the fixation duration d_i. For each centroid, the weighted distance is then summed for all fixations to form a vector of size equal to the number of ROI.

7.6.2 Incremental classification

We perform supervised learning to classify the extracted features into the respective attributes (demographic, personality, interests) for every shot. As discussed in Chapter 7.6, for any single shot, the classification results are likely to be poor for some attributes and better for others. Instead of returning the mixed results, we can exploit the temporal ordering of the shots to incrementally improve the results by combining the classification results of the same attribute from previous shots from the same video. To this end, we implemented a supervised meta-classifier which treats the ordered set of shot classification results as the input features. The size of feature vector is equal to the current shot index.

Just like a conventional tracking profiling method follows a user from one website to another, improving on the accuracy of the profile; our incre-

mental classifier leverages on the past classification results for better final classification. It is interesting to note that the similarities between the two methods, except that our time frame is significantly shorter: seconds vs days.

7.7 Empirical experiments

To the best of our knowledge, we are the first to propose an instant implicit user profiling system. As such, there are many interesting and valid research questions. We conducted empirical experiments to answer two important questions about our user profiling method.

Q1. With respect to targeted advertising in videos, can acceptably accurate profiles be inferred with our proposed method?

Q2. How fast can the acceptably accurate profile be consistently returned?

Q1 is critical as the system is useless if the inferred profiles are inaccurate. We choose to use *accuracy* instead of other statistical scores such *F1* measure as the attributes are quite well-balanced and the outcomes of the binary classification are equally important (See Table 7.4). The answer to Q2 will justify our claim that our proposed method can *accurately* and *consistently* profile users in a timely manners.

7.7.1 Data preparation

We refer to personal traits (e.g. gender, age, personality types) and topics of interest (e.g. animals, computers) collectively as `attributes`.

For the presentation of the experimental results, the attributes are abbreviated as: field of study/work ⇒ *specialty*, highest education qualifications ⇒ *education*, personal income ⇒ *personal*, household income ⇒ *household*; extrovert/introvert ⇒ *ei*; sensing/intuition ⇒ *sn* and think-

TRAITS	MAJORITY	MINORITY
Gender	female(0.59)	male(0.41)
Agegroup	$\leq$24(0.76)	$\geq$25(0.24)
Ethnicity	chinese(0.69)	others(0.31)
Religiosity	religious(0.67)	none(0.33)
Specialty	sci&eng(0.65)	others(0.35)
Education	tertiary(0.69)	post-grad(0.31)
Income	0-999(0.71)	$\geq$1000(0.29)
Household	1-4999(0.75)	$\geq$5000(0.25)
EI	Introvert(0.53)	Extrovert(0.47)
SN	Sensing(0.57)	Intuition(0.43)
TF	Feeling(0.63)	Thinking(0.37)

Table 7.4: Grouping of traits for the dataset. The numbers in parentheses show the distribution of the traits.

ing/feeling $\Rightarrow tf$.

The recorded eye-gaze data were preprocessed by the vendor's software to extract the fixations. Fixations from the preferred eye as indicated by the subjects were used. In a live system, the vendor's software is able to return fixation data in real-time. Missing and noisy eye-tracking data was not removed from our experiments to simulate actual live systems.

For the inference of topics of interest, only 1 participant indicated interests in *real* estate. Hence, this topic is removed for consideration, leaving 26 topics of interest.

Each video is manually segmented into shots. The number of shots are: 107, 155, 135 and 140 respectively. We briefly considered performing classification using the whole set of eye-gaze data for a single video. However, the eye-gaze over the entire video is too diverse and will not be useful for our purpose. On the other extreme, eye-gaze data on a single frame is insufficient for classification. Instead, shot segmentation allows eye-gaze data on each set of content-coherent and semantic-similar frames to be classified separately.

7.7.2 Experimental results

The objective of the experiments is to classify each attribute (trait or topic of interest) into 2 possible classes. For topics of interest, the 2 classes are "interested" and "not-interested". For traits with multiple possible values, we consolidate them into 2 groups for a more even distribution. See Table 7.4 for the groupings of traits and the distributions of the population (*Prior* is ratio of the majority class).

For each shot in each video, the 2 feature vectors: *Stats* and *ROI* are extracted from the fixations of each person. A support vector machine (SVM) classifier is trained per shot per attribute. We used the standard linear SVM classifier in the Matlab Biometric Toolbox, with the default parameters and auto-scaling.

Using incremental classification method, the ordered classification results from the previous and current shots form the input feature vector for the meta-classifier, also a SVM (same implementation and parameters as the per-shot classifiers). Leave-one-out cross validation is used to evaluate the meta-classifiers, i.e. a single subject is left out of the training set in each round. Figure 7.3 shows the example of classifying *gender* trait for *satire* video with the *Stats* features.

Acceptable mean accuracy

How well does `Eye-2-I` work? As we are using supervised training, the *baseline* accuracy should be *Prior*, the fraction of the majority class in the population. Hereafter, we refer to accuracies which are higher than the respective *Prior* as *acceptable*.

The mean and peak accuracies from *Stats* and *ROI* are plotted in Figure 7.4 for traits and in Figure 7.5 for topics of interest. The statistics of the accuracies are computed from the set of leave-one-subject-out accuracy of every shot from the 4 videos.

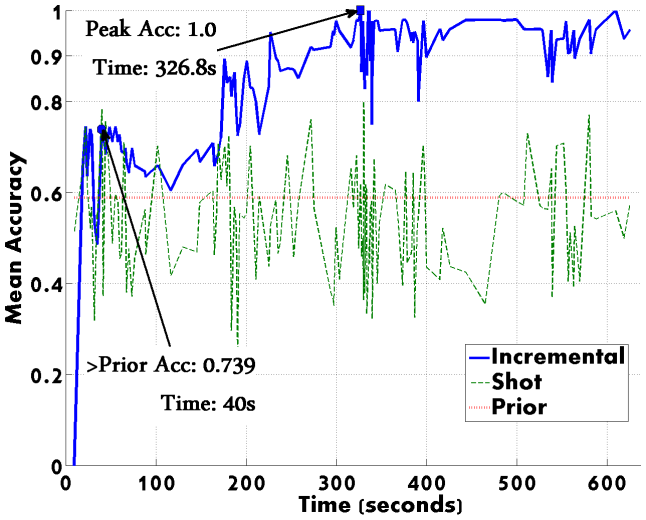

Figure 7.3: Mean Accuracy vs Time plot for *gender* trait classification (*Stats* feature) with *satire* video. Incremental classifier's accuracies improve over time. It achieves consistently better than *Prior* accuracy at 40 seconds. It peaks at perfect accuracy after 326.8 seconds (5.4 minutes). After that time, with a few exceptions, accuracy of > 0.9 is sustained. Single-shot classifiers' accuracies depend only on the respective shots, and perform much worse, especially towards end of the video. This applies to all attributes with every video.

Our proposed incremental classification method is able to *acceptably* classify 26 attributes using *Stats*; and 29 attributes with *ROI*, out of the 37 attributes. Furthermore, the peak accuracies are quite remarkable, achieving more than 0.9 for every attribute; and obtaining perfect classification for many attributes, with either feature.

From the experimental results, the answer to Q1 is that our proposed method is *acceptably* accurate profiling tool for a non-critical application like targeted advertisement. Moderately high mean accuracies are obtained for most attributes, the peak accuracies are also very high. One important advantage of our method is that the accuracies increase as more eye-gaze data becomes available.

Stats vs ROI

From Figure 7.4 and 7.5, *Stats* has *acceptable* mean accuracies for 7 out of 8 demographic traits, all 3 personality types and 16 out of 26 topics

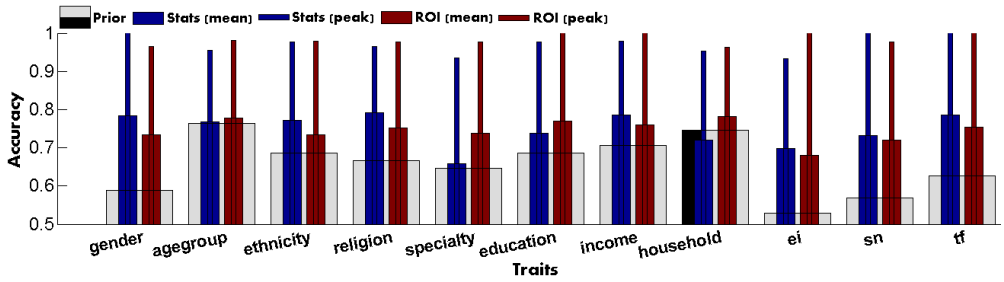

Figure 7.4: Bar chart comparing mean and peak accuracy of *Stats* and *ROI* against *Prior* for incremental classification of personal traits. Only *household* trait with *Stats* is lower than *Prior* (shaded in black). *ROI* outperforms *Stats* for the following traits: *agegroup*, *specialty*, *education* and *household*. The peak accuracies are above 0.9 for both features across all traits. There are several perfect peak classifications.

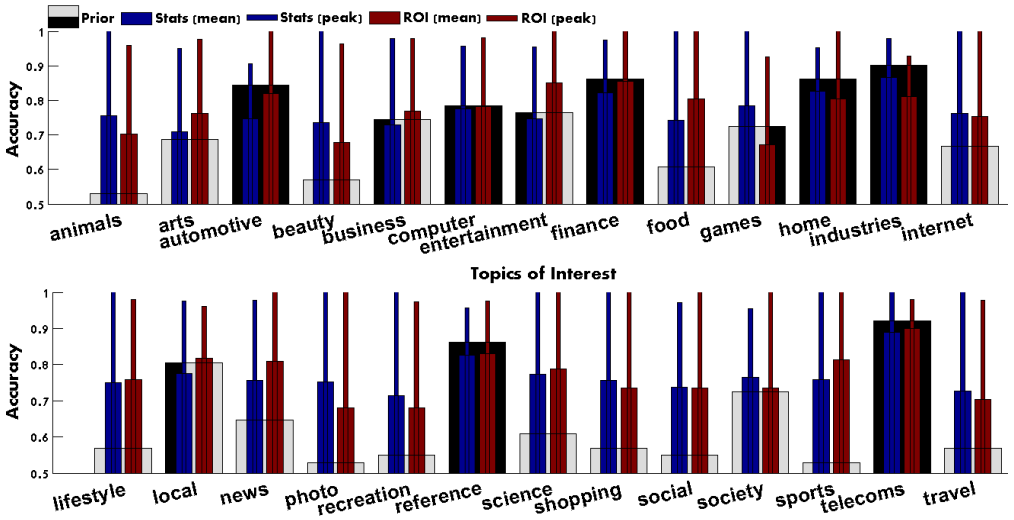

Figure 7.5: Bar chart comparing mean and peak accuracy of *Stats* and *ROI* against *Prior* for incremental classifications of topics of interest. Only 7 (shaded in black) out 26 topics have mean accuracies lower than *Prior* for both features. The peak accuracies are above 0.9 for both features across all topics. The lowest peak accuracy is for *Automotive* topic (*Stats*=0.907).

of interest. *ROI* has *acceptable* accuracies for all 8 demographic traits, all 3 personality types and 18 out of 26 topics of interest. Overall, *ROI* can classify 3 more attributes with *acceptable* accuracies than *Stats*. This makes *ROI* a better feature than *Stats*.

Comparing the mean accuracies, *Stats* is more accurate for traits inference (7 out of 11). This is in agreement with various prior studies that statistics of the eye-gaze are correlated to some personal traits Goldstein et al. (2007); Chua et al. (2005); Wu et al. (2013). *ROI* outperforms *Stats*

		50%		70%	
		STATS	**ROI**	**STATS**	**ROI**
TRAITS	*Documentary*	5:53	5:03	9:18	5:34
	Animation	4:56	3:24	6:18	6:18
	Satire	5:35	5:40	5:37	6:18
	Romance	2:16	5:10	3:49	7:33
INTERESTS	*Documentary*	5:30	5:50	6:49	8:46
	Animation	6:18	6:18	8:36	7:01
	Satire	5:37	5:40	9:40	8:59
	Romance	6:11	3:50	10:11	9:41

Table 7.5: Time in minutes taken to *consistently* and *accurately* classify 50% and 70% of the traits and topics of interests respectively. The faster time between the 2 features are underlined.

for inferring interests (14 out of 26 topics for mean accuracies).

7.7.3 Response time for consistently accurate classification

With our incremental classifier, the accuracy will improve as more eye-gaze data becomes available. When will *acceptably* accurate profiles be *consistently* available?

Table 7.5 shows the time taken for each video to *acceptably* classify at least 6 and 8 traits *consistently*, that is 50% and 70% of the 11 traits. The *romance* video is the fastest. With it, 8 traits (70%) can be *consistently* and *acceptably* classified by 3:49 minutes (*Stats*).

For all 4 videos, the classifiers are able to *consistently* and acceptably infer 6 traits by 5:35 minutes; and 8 traits by 6:18 minutes.

Also from Table 7.5, we measure the time taken for *acceptably* infering at least 13 (50%) and 18 (70%) of the 26 topics *consistently*. For all 4 videos, the classifiers are able to *consistently* and *acceptably* infer 13 topics by 6:18 minutes; and 18 topics by 9:41 minutes.

To answer Q2., for any video, it takes *at most 6:18 minutes* to *consistently* and *acceptably* infer 50% of the attributes; it takes at most *9:41 minutes* to *consistently* and *acceptably* infer 70% of the attributes. With the *romance* video, the response time is only 3:50 minutes for 8 traits and 13 topics. These are reasonable response time for targeted advertisements as too early insertion of advertisements may be disruptive to the viewing experience.

7.8 Discussions

Our experimental results show that eye-gaze can be used to accurately and quickly profile a user implicitly for all of the 4 videos. The video should be more than 6 minutes for consistently accurate profiling. Out of the 26 topics of interest, only 7 topics: *automotive*, *computer*, $finance$, *home*, *industries*, $reference$, and *telecommunications*; have *unacceptable* mean accuracies for both features. We hypothesize the low accuracies to the fact that these topics are not related to the videos. More experiments would need to be performed to ascertain this hypothesis.

While there are many applications which require a detailed accurate profile within seconds, no method can achieve this. Questionnaires and web-tracking need minutes and days respectively. Appearance methods such as faces, while very fast, are limited to very few attributes such as gender, age etc. But *Eye-2-I* is able to provide a detailed profile (lower accuracies) by the first video shot (about 2 to 20 seconds). This is a clear advantage. For higher accuracy, it may require 4 or 5 minutes of eye-tracking data. But it is still much faster compared to web tracking (days/weeks). Compared to questionnaires, *Eye-2-I* is implicit.

An interesting scientific question to ask is what are the visual or semantic features which can determine if a visual stimulus is more suitable

for classification of an attribute, e.g. *gender*. The answer to this question demands contributions from multiple disciplines such as behavioral psychology, computer science and even neuro-psychology. Our experimental results provide some hints. We observe that videos which are more affective, that is *romance* and *animation* are better than the *documentary* video for traits classification. Also that topics of interests which are more related to the video tends to have better accuracies. By using incremental classification for the video-watching task, the choice of the visual stimuli becomes much less critical.

A limitation in our current setup for `Eye-2-I` is the requirement for sufficient labeled data for each video. This can be overcome using crowdsourcing. Each video is initialized with a profile of the expected population distributions, e.g. 0.5 male, 0.5 female. After watching a video, a user will be prompted to update his/her inferred profile. A suitable classification algorithm will use the newly labeled data for **online learning**, after that the labeled data can be safely discarded. Online machine learning is a model of induction that learns one instance at a time. The goal in online learning is to predict labels for instances. The key defining characteristic of on-line learning is that soon after the prediction is made, the true label of the instance is discovered. This information can then be used to refine the prediction hypothesis used by the algorithm. The goal of the algorithm is to make predictions that are close to the true labels. As more labeled data becomes available, the system's accuracy will improve.

We acknowledge that our sample size of 51 participants and 4 videos is too small to draw a definite conclusion that every attribute can be accurately inferred in the general population with our method. With additional resources, this interesting problem further by recruiting more participants from the general population, especially seniors and children.

Another open research question which we hope to address in the future

is the stability and repeatability of our approach. Can `Eye-2-I` accurately infer profile from a user who had watched the video before? Our dataset's protocol is not designed to answer this question and a new dataset will need to be collected.

Our work will prove most useful to people who are interested in user profiling but are limited by the several drawbacks of the current methods. In particular, we conjecture that our proposed method is ideal for systems which instant user profiling are highly desirable, but are impossible with current methods: e.g. public kiosks. Our method can also be used to complement current methods to obtain higher accuracies or when some attributes are not easily inferred by current methods, e.g. interests.

Chapter 8

Advancing Computational Eye-gaze Research

```
If I have seen further
it is by standing on the shoulders of Giants.

- Isaac Newton
```

Our proposed framework is the first to formalize the VIP variables which influences eye-gaze. With this formal framework as a reference, the features, computational models and assumptions of applications and research problems can be formally described and compared in the form of Equation 3.6. New research directions can also be easier to discover by identifying gaps of existing models. The assumptions made by existing applications and research can be systematically reviewed.

As an analogy, our VIP model can be compared against the periodic table of chemical elements (Mendeľčev, 1895). It is used to guide the discovery of new elements (research/applications) and to understand and categorize the properties of existing chemical elements. Furthermore, it can be used to extrapolate properties of undiscovered elements. Therefore, our VIP framework is a useful tool for the research community to further advance

the exciting field of computational eye-gaze research.

Tables 8.1 and 8.2 list the various eye-gaze applications and their formulation with reference to the VIP framework. From these tables, we can identify improvements for our trait inference application by using the more advanced mRMR feature selection algorithm as implemented by the activity recognition systems (Bulling et al., 2011). The inclusion of saccades into our trait inference application should also hold much promise. We will present more examples of the usefulness of our framework for advancing the computational eye-gaze research in the rest of this chapter.

Application	Formulation	f/f^{-1}	E	V	I	P
Saliency (Bruce and Tsotsos, 2006)	$E = \underset{I=c}{f}(V)$	Information Maximization	Gaussian Filter of Fixations	Colors	Free-view	–
Saliency (Judd et al., 2009)	$E = \underset{I=c}{f}(V)$	SVM	Gaussian Filter of Fixations	33 Features: Energy, Subband, Intensity, Orientation, Color, Intensity, Horizon, Face, Person, Center Bias, etc.	Free-view	–
Saliency (Ouerhani et al., 2004)	$E = \underset{I=c}{f}(V)$	Conspicuity, Normalization, Summation	Gaussian Filter of Fixations	Colors, Orientation	Free-view	–
Saliency (Le Meur et al., 2006)	$E = \underset{I=c}{f}(V)$	Masking, Interactions, Grouping	Gaussian Filter of Fixations	Colors, Intensity	Free-view	–
Gaze Modeling (Steptoe et al., 2009)	$E = f(V)$	Random	Fixations, Saccades	Faces, Objects	–	–
Image Segmentation (Ramanathan et al., 2010)	$V = \underset{I=c}{f^{-1}}(E)$	Energy Minimization	Fixations	Most Salient Region	Free-view	–
Object Labeling (Ramanathan et al., 2009)	$V = \underset{I=c}{f^{-1}}(E)$	Clustering, Color Segmentation	Fixations, Saccades	Affective Objects	Free-view	–
Video Summarization (Vural and Akgul, 2009)	$V = \underset{I=c}{f^{-1}}(E)$	Energy Minimization	Gaze	Motions	Surveillance	–
Activity Recognition (Bulling et al., 2011)	$I = f^{-1}(E)$	mRMR, SVM	Saccades, Fixations, Blinks	–	Activity	–
Midas Touch (Bednarik et al., 2012)	$I = f^{-1}(E)$	Normalization, SVM	Fixation	–	Act/ Observe	–
Attention State (Asteriadis et al., 2009)	$I = f^{-1}(E)$	Fuzzy Inference	Gaze	–	Attention	–
Biometric (Kinnunen et al., 2010)	$P = f^{-1}(E)$	UBM, GMM	Gaze	–	–	Identity
Twins Identification (Zhang et al., 2013)	$P = \underset{I=c}{f}^{-1}(E)$	Alignment, GMM, SVM	Gaze, Pupil Movement, Open Magnitude	–	Talking	Identity
Video Summarization (Katti et al., 2011)	$I = f^{-1}(E, V)$	Threshold	Pupillary Dilations	Shot Boundary	Emotion	–
Interactive Ads (Yadati et al., 2013)	$I = f^{-1}(E, V)$	Fusion	Pupillary Dilations	Affect	Emotion	–
Implicit Tagging (Gao et al., 2009)	$I \approx f^{-1}(E, V)$	AIC, HITV	Pupillary Dilations	Intensity	Emotion	–
Smart Pause (Samsung, 2013)	$I \approx f^{-1}(E, V)$	Proprietary	Gaze	Video/ Other	Pause/ Play/Other	–

Table 8.1: Some examples of applications and their corresponding VIP models. For brevity, we slightly modify the conventional meanings of "=" and "f/f^{-1}". "=" means that the objective is to minimize the error between the left and right side of the equation. The error measure is as per defined by the respective papers. "f^{-1}" means the inverse dependency of eye-gaze and the VIP factors. There is no implication that a corresponding "f" must exist.

Application	Formulation	f/f^{-1}	E	V	I	P
Trait Inference	$P = f^{-1}_{V=c1,I=c2}(E)$	Correlation, SVM	Fixations	Specific Images	Free-View	Demography, Personality
Trait-Specific Fixation Prediction	$E = f_{I=c}(V,P)$	Conditional Gaussian Filter	Fixations	Prior Fixations	Free-View	Demography, Personality
Eye-2-I	$(I,P) = f^{-1}(E,V)$	Incremental classification (SVM)	Pupillary Dilations, Fixations	Audio, Visual	Arousal, Interests	Demography, Personality

Table 8.2: Our applications and their VIP features.

8.1 Gaps in existing areas

8.1.1 Computational Human Visual Attention

The primary variables in computational HVA research are the “bottom-up" cues and the “top-down" influences. The “bottom-up” cues are subsumed by our proposed V variable and the “top-down” influences are subsumed in the I variable. From the various HVA datasets, there are non-trivial differences of fixations between subjects which cannot be explained by the current top-down/bottom-up model (Judd et al., 2012). Our VIP model suggests that these differences are due to the P variable. Evidences from both psychology studies (Goldstein et al., 2007; Chua et al., 2005; Risko et al., 2011) and our experimental results in Chapter 6 validate this.

8.1.2 Vision and Multimedia

Similarly, the computer vision and multimedia researchers relied on the “bottom-up/top-down” model. And therefore, it suffers from the same limitations of computational HVA. We have developed a new application, trait inference from eye-gaze which uses the third variable, P. Furthermore, we propose a new implicit just-in-profiling system in Chapter 7. Other existing applications will likely to benefit from modeling the complete VIP variables.

8.1.3 Biometrics

The biometrics researchers are concerned with using the eye-gaze data to derive the identity of the viewer, hence their models are subsumed by the P variable in our framework. Current, biometrics models are P-only. One potential new research would be the IP model. For example, by using pupillary dilation, a security systems may be able to identify an authorized user who is in an abnormal emotional state.

8.1.4 Human Computer Interface

In user-interface research, the reference model is one of the VI. There is no consideration of the P factors. As far as our extensive literature review shows, in all existing models, only one or two of the factors are considered. These models are incomplete without representing the personal traits of the user. For example, there are age, gender and cultural differences in eye-gaze when conversing (Argyle and Cook, 1976). Therefore, a male young Japanese avatar with a female elderly Canadian eye-gaze model may impede the level of immersion in the virtual environment.

8.2 Comparisons

Our framework allows direct comparisons between research from different fields.

8.2.1 HVA vs Biometric

In computational HVA, the perfect computational model is one which can *exactly* and *completely* predict the visual attention (therefore, eye-gaze) of *any* person, given *any* bottom-up/top-down features. From evidences of empirical data about differences of eye-gaze among individuals, this model

is unlikely to exist. However, it has not been formally proven its impossibility.

Our informal proof is as follows. For a particular eye-gaze biometric system, such as (Holland and Komogortsev, 2011), individuals can be identified by their eye-gaze data (fixations, saccades, scanpaths) with certain specific task (reading) and stimulus (excerpt from Lewis Carroll's "The Hunting of the Snark"). But given the same task and stimulus, the perfect computational HVA model will predict only one set of eye-gaze data. This is clearly a contradiction since it is impossible to identify the different individuals with the same input eye-gaze data.

Our formal proof:

1. For biometric, $P_b = \underset{V_b=c1,I_b=c2}{f_b^{-1}}(E_b)$ where P_b is the identities of the viewers.
2. For HVA, $E_{HVA} = f_{HVA}(V_{HVA}, I_{HVA})$
3. Let $V_{HVA} = c1$, $I_{HVA} = c2$ and $E_b = E_{HVA}$, then $P_b = \underset{V_b=c1,I_b=c2}{f_b^{-1}}(f_{HVA}(c1, c2))$.
4. Since $f_{HVA}(c1, c2)$ is a constant, then P_b is a constant. A contradiction.

In summary, if the stimulus (V) and the task (P) can be used to identify different viewers from their eye-gaze E, then no VI-type saliency prediction algorithm will be able to exactly predict the eye-gaze (E) from the same V and P. If a VI saliency prediction algorithm can exactly predict the eye-gaze from V and I, then no eye-gaze biometric can identify the viewers. Our framework concisely and formally describes the contrasting goals of the two systems.

8.2.2 P for Privacy

Our experimental results in Chapter 5 show that various personal traits can be reliably inferred from eye-gaze with some specific stimuli. This has

important implications for other applications. For example, when designing a biometric eye-gaze or an eye-gaze input systems, it is important to consider if the stimulus chosen can unintentionally reveal sensitive personal traits about the users. Without our framework, the implications of our trait inference problem to other applications are less obvious.

To mitigate the risks of compromising privacy, we proposed the `Eye-2-I` system in Chapter 7. The inferred profiles can be stored locally and transiently as the inference can be performed just-in-time.

Chapter 9

Conclusion

```
The gaze of man is free to move around
From place to place where'ere the eye does will.
It flicks about to give the mind its fill
And make the image whole within the head.
It seeks with lightning speed the source of sound
And follows smoothly anywhere it's led.

From deep within the brain the signals come
To stablize the world of visual space
Against all violent motion of the face:
And does it all with simple rules of thumb.

- John W. Senders (Professor of Everything)
```

In conclusion, eye-gaze data holds many promises as it bridges the semantic gap between the low-level visual features and the high-level abstract concepts. The recent advances in hardware, psychology research and computer applications also build to an crescendo unparalleled in the history for this modality to be seriously regarded by computer scientists. For our contributions, we proposed a novel VIP framework which unifies all current

eye-gaze computational models. This framework will facilitate the advances of eye-gaze research as new problems can be more easily identified. Secondly, we identified and solved two new research problems: personal traits inference and developed an implicit just-in-time profiling system: `Eye-2-I`. Thirdly, we proposed a trait-specific fixation prediction approach which outperforms current trait-agnostic approach for some images.

In our view, our contribution allows researchers to have a broader perspective of factors which affect eye-gaze and their implications. Thus, in view of the intellectual challenges posed and its tremendous promises, eye-gaze data should be recognized as a critical modality in computer science research, development, systems, and applications. We humbly stand on the collective shoulders of many hard-working researchers from diverse disciplines to propose the unifying VIP framework for eye-gaze research. An exciting world with many new possibilities awaits when so much can be known just from one's eye-gaze.

My thesis is the end of my PhD voyage of discovery. It is also the beginning of incorporating Personal traits into computational eye-gaze research. As a new research direction, there is clearly much to be done. I list some of these below:

1. Investigate the influence of the 4th factor: Environment While our VIP framework is complete in the controlled environment, the external environment is an important factor for deployment in live systems. Environmental factors such as lightings, ambient noise etc are likely to have influence on eye-gaze.
2. Explore other machine learning techniques Our current applications require sufficient labeled data for each visual stimulus (image/video) as we using supervised learning methods. A more generalized approach is to leverage on online learning and transfer learning approaches. Deep-learning methods have been used successfully for

many challenging computer vision problems such as large-scale object classification and face-recognition. These methods may also be useful for our problems.

3. Explore new features, e.g. scanpath, for our applications There are other eye-gaze features which may be useful for our applications, e.g. scanpath and saccades.
4. Examine the interaction between different factors While our VIP framework is general, our applications assume that the various traits, interests and emotions are independent. Our datasets are too small to conclusively investigate the interactions between these factors.
5. Effects of repeated viewing Our experimental results are based on the first viewing. However, we noted that repeated viewing of the same stimulus is an important I factor, as prior knowledge is known to be influenced eye-gaze. Future work may be extended to include this.
6. More user study Due to the tremendous amount of resources required for the acquisition of eye-gaze dataset, we are limited to a small population of 72 and 51 subjects for VIP and VVIP respectively. The number of images are 150 for VIP; and 4 videos for VVIP. These number while among the highest in eye-gaze datasets (Winkler and Subramanian, 2013), may not be sufficient for investigating other topics: interactions between different traits; transfer learning etc.
7. Include other traits/interests There are other personal traits and interests which are not included in our studies and may be important for some applications. Some examples are hobby and political views.

It is, on one hand, regrettable that due to resource constraints, the list is longer than what I would have liked. On the other hand, being a totally new research framework, it is exciting that there is so much more to discover.

Finally, while my PhD voyage is coming to an end; another voyage,

more challenging, more rewarding, more discoveries to be made, begins.

Bibliography

Adobe Systems and Edelman Berland (2012). Click here: The state of online advertising. http://www.adobe.com/aboutadobe/pressroom/pdfs/Adobe_State_of_Online_Advertising_Study.pdf.

Amazon (2014). Amazon Fire Phone - 13MP Camera, 32GB - Shop Now. http://www.amazon.com/gp/product/B00E0E0WKQ. Accessed: 26/09/2014.

Arbeláez, P. and Cohen, L. (2008). Constrained image segmentation from hierarchical boundaries. In *CVPR 2008*, pages 1–8. IEEE.

Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze.* Cambridge U Press.

Arnon Amir, C. C., Myron Dale Flickner, S. J. C., David Bruce Koons, S. J. C., and Gregory Fraser Russell, Y. H. N. (2003). Calibration-free eye gaze tracking. Patent. US 6578962.

Asteriadis, S., Tzouveli, P., Karpouzis, K., and Kollias, S. (2009). Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. *Multimedia Tools and Applications*, 41(3):469–493.

Bagon, S., Boiman, O., and Irani, M. (2008). What is a good image segment? a unified approach to segment extraction. In *ECCV 2008*, pages 30–44. Springer.

Barber, P. J. and Legge, D. (1976). *Psychological types*, chapter 4: Information Acquistion. Methuen, London, UK.

Bednarik, R., Kinnunen, T., Mihaila, A., and Fränti, P. (2005). Eye-movements as a biometric. *Image analysis*, pages 16–26.

Bednarik, R., Vrzakova, H., and Hradis, M. (2012). What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 83–90. ACM.

Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207.

Borji, A., Sihite, D. N., and Itti, L. (2013a). Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *Image Processing, IEEE Transactions on*, 22(1):55–69.

Borji, A., Tavakoli, H. R., Sihite, D. N., and Itti, L. (2013b). Analysis of scores, datasets, and models in visual saliency prediction. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 921–928. IEEE.

Bradley, M. M., Miccoli, L., Escrig, M. A., and Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

Bressan, P., Kramer, P., and Germani, M. (2008). Visual attentional capture predicts belief in a meaningful world. *Cortex*, 44(10):1299–1306.

Briggs, I. and Myers, P. B. (1980). *Gifts differing: Understanding personality type*. Davies-Black Publishing.

Bruce, N. and Tsotsos, J. (2006). Saliency based on information maximization. *Advances in neural information processing systems*, 18:155.

Buchala, S., Davey, N., Gale, T. M., and Frank, R. J. (2005). Principal component analysis of gender, ethnicity, age, and identity of face images. *Proc. IEEE ICMI*.

Bulling, A., Ward, J., Gellersen, H., and Troster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *Pattern Analysis and Machine Intelligence*, 33(4):741–753.

Chen, Y., Nguyen, T. V., Kankanhalli, M., Yuan, J., Yan, S., and Wang, M. (2014). Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chua, H., Boland, J., and Nisbett, R. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12629–12633.

CodeProject (2013). TrackEye : Real-Time Tracking Of Human Eyes Using a Webcam - CodeProject. `http://www.codeproject.com/Articles/26897/TrackEye-Real-Time-Tracking-Of-Human-Eyes-Using-a`. Accessed: 02/04/2013.

Coppersmith, D., Hong, S. J., and Hosking, J. R. (1999). Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197–217.

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215.

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4):455–470.

Elazary, L. and Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3).

emedia (2007). Social networking sites: Almost two thirds of users enter false information to protect identity. `http://www.realwire.com/release_detail.asp?ReleaseID=6671`. Accessed: 14/03/2014.

Facebook (2014). Create an account. `https://www.facebook.com/help/www/345121355559712`. Accessed: 14/03/2014.

Frintrop, S., Rome, E., and Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6.

Gao, Y., Barreto, A., and Adjouadi, M. (2009). Monitoring and processing of the pupil diameter signal for affective assessment of a computer user. In *Human-Computer Interaction. New Trends*, pages 49–58. Springer.

Gevers, T. (2014). eMotion emotion analyzer. `http://http://visual-recognition.nl/`.

Gibbs, G. (2013). 38 Million Adobe Users Affected by Security Breach. `http://www.girardgibbs.com/adobe-data-breach-`. Accessed: 14/03/2014.

Goldstein, R., Woods, R., and Peli, E. (2007). Where people look when watching movies: Do all viewers look at the same place? *Computers in biology and medicine*, 37(7):957–964.

Google (2014a). How ads are targeted to your site. `https://support.google.com/adsense/answer/9713?hl=en`. Accessed: 14/03/2014.

Google (2014b). How Google infers interest and demographic categories. https://support.google.com/adsense/answer/140378?hl=en&ref_topic=23402. Accessed: 14/03/2014.

Gunes, H. and Piccardi, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):64–84.

Hanjalic, A. and Xu, L.-Q. (2005). Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154.

Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500.

Hirsh, J. B., Kang, S. K., and Bodenhausen, G. V. (2012). Personalized persuasion tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6):578–581.

Hodge, A. and Rosenblatt, M. (2013). Electronic devices with gaze detection capabilities. US Patent 20,130,135,198.

Hoffman, J. E. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Attention, Perception, & Psychophysics*, 57(6):787–795.

Holland, C. and Komogortsev, O. V. (2011). Biometric identification via eye movement scanpaths in reading. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE.

Hou, X., Harel, J., and Koch, C. (2012). Image signature: Highlighting sparse salient regions. *Pattern Analysis and Machine Intelligence*, 34(1):194–201.

Hunziker, H. W. (1970). Visual information reception and intelligence: An investigation of the role of eye movements in problem solving. *Psychologie v Ekonomicke Praxi*, 29:165–171.

Hunziker, H.-W. (2006). Im auge des lesers: foveale und periphere wahrnehmung-vom buchstabieren zur lesefreude. *the eye of the reader: foveal and peripheral perception-from letter recognition to the joy of reading), Transmedia Zurich.*

Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3):4.

Jaimes, A., Pelz, J. B., Grabowski, T., Babcock, J. S., and Chang, S.-F. (2001). Using human observer eye movements in automatic image classifiers. In *Photonics West 2001-Electronic Imaging*, pages 373–384. International Society for Optics and Photonics.

Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. Technical report, MIT.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV).*

Jung, C. G., Baynes, H., and Hull, R. (1991). *Psychological types.* Routledge London, UK.

Katti, H., Yadati, K., Kankanhalli, M., and Tat-Seng, C. (2011). Affective video summarization and story board generation using pupillary dilation and eye gaze. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 319–326. IEEE.

Kawamoto, D. and Mills, E. (2006). AOL apologizes for release of user search data. http://news.cnet.com/AOL-apologizes-for-release-of-user-search-data/2100-1030_3-6102793.html. Accessed: 14/03/2014.

Kinnunen, T., Sedlak, F., and Bednarik, R. (2010). Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 187–190. ACM.

Koelstra, S. and Patras, I. (2013). Fusion of facial expressions and {EEG} for implicit affective tagging. *Image and Vision Computing*, 31(2):164 – 174. Affect Analysis In Continuous Input.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*.

Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., and Yan, S. (2012). Depth matters: influence of depth cues on visual saliency. In *Computer Vision–ECCV 2012*, pages 101–115. Springer.

Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence*, 28(5):802–817.

Liao, W.-S., Chen, K.-T., and Hsu, W. H. (2008). Adimage: Video advertising by image matching and ad scheduling optimization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 767–768. ACM.

Ma, K.-T., Sim, T., and Kankanhalli, M. (2013). VIP: A unifying framework for computational eye-gaze research. In *4th International Workshop on Human Behavior Understanding*. Springer.

Martens, L. (2012). *Automatic Person and Personality Recognition from Facial Expressions*. PhD thesis, Tilburg University.

Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3):229–240.

Mei, T., Hua, X.-S., Yang, L., and Li, S. (2007). Videosense: towards effective online video advertising. In *Proceedings of the 15th International Conference on Multimedia*, pages 1075–1084. ACM.

Mendeľěev, D. (1895). *Ueber die Beziehungen der Eigenschaften zu den Atomgewichten der Elemente*.

Mishra, A., Aloimonos, Y., and Fah, C. L. (2009). Active segmentation with fixation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 468–475. IEEE.

Miyahira, A., Morita, K., Yamaguchi, H., Morita, Y., and Maeda, H. (2001). Gender differences and reproducibility in exploratory eye movements of normal subjects. *Psychiatry and clinical neurosciences*, 54(1):31–36.

Ni, B., Xu, M., Nguyen, T., Wang, M., Lang, C., Huang, Z., and Yan, S. (2014). Touch saliency: Characteristics and prediction. *Transactions on Multimedia*, 16(6):1779 – 1791.

Ouerhani, N., Von Wartburg, R., Hugli, H., and Muri, R. (2004). Empirical validation of the saliency-based model of visual attention. *Electronic letters on computer vision and image analysis*, 3(1):13–24.

Pal, R., Mukherjee, J., and Mitra, P. (2009). An approach for preparing groundtruth data and evaluating visual saliency models. In *Pattern Recognition and Machine Intelligence*, pages 279–284. Springer.

Pantic, M. and Vinciarelli, A. (2009). Implicit human-centered tagging [social sciences]. *Signal Processing Magazine, IEEE*, 26(6):173–180.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718.

Rajaraman, A. and Ullman, J. D. (2012). *Mining of massive datasets*. Cambridge University Press.

Ramanathan, S., Katti, H., Huang, R., Chua, T.-S., and Kankanhalli, M. (2009). Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 729–732. ACM.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., and Chua, T.-S. (2010). An eye fixation database for saliency detection in images. In *Computer Vision–ECCV 2010*, pages 30–43. Springer.

Rigas, I., Economou, G., and Fotopoulos, S. (2012). Human eye movements as a trait for biometrical identification. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 217–222. IEEE.

Risko, E. F., Anderson, N. C., Lanthier, S., and Kingstone, A. (2011). Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*.

Rosenberg, L. (2006). Gaze-responsive video advertisment display. US Patent App. 11/465,777.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Samsung (2013). Samsung Galaxy S4 - Life Task. http://www.samsung.com/global/microsite/galaxys4/lifetask.html#page=pausescroll. Accessed: 02/04/2013.

Schleicher, R., Galley, N., Briest, S., and Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010.

SensoMotoric Instruments (2013). SensoMotoric Instruments GmbH > Gaze and Eye Tracking Systems > Products > iView X Hi-Speed. http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/iview-x-hi-speed.html. Accessed: 02/04/2013.

Shen, C. and Zhao, Q. (2014). Learning to predict eye fixations for semantic contents using multi-layer sparse network. *Neurocomputing*, 138:61–68.

Shen, J. and Itti, L. (2012). Top-down influences on visual attention during listening are modulated by observer sex. *Vision research*, 65:62–76.

SMI (2014). Smi eye tracking glasses, mobile eye tracking glasses by sensomotoric instruments (smi). http://eyetracking-glasses.com/. Accessed: 26/09/2014.

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3:42–55. Issue 1.

Song, G., Pellerin, D., Granjon, L., et al. (2013). Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research*, 6(4):1–13.

Steptoe, W., Oyekoya, O., Murgia, A., Wolff, R., Rae, J., Guimaraes, E., Roberts, D., and Steed, A. (2009). Eye tracking for avatar eye gaze

control during object-focused multiparty interaction in immersive collaborative virtual environments. In *Virtual Reality Conference, 2009. VR 2009. IEEE*, pages 83–90. IEEE.

Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *Systems, Man and Cybernetics, IEEE Transactions on*, (1):116–132.

Tobii (2013). Tobii REX brings Gaze eye-tracking tech to any Windows 8 machine. `http://www.engadget.com/2013/01/02/tobii-rex/`. Accessed: 19/04/2013.

Vural, U. and Akgul, Y. S. (2009). Eye-gaze based real-time surveillance video synopsis. *Pattern Recognition Letters*, 30(12):1151–1159.

Winkler, S. and Subramanian, R. (2013). Overview of eye tracking datasets. In *Workshop on Quality of Multimedia Experience*.

Wu, D. W.-L., Bischof, W. F., Anderson, N. C., Jakobsen, T., and Kingstone, A. (2013). The influence of personality on social attention. *Personality and Individual Differences*.

Yadati, K., Katti, H., and Kankanhalli, M. (2013). Interactive video advertising: A multimodal affective approach. In *Advances in Multimedia Modeling*, pages 106–117. Springer.

Yadati, K., Katti, H., and Kankanhalli, M. (2014). CAVVA: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 16(1).

Yankelovich, D. and Meer, D. (2006). Rediscovering market segmentation. *Harvard Business Review*, 84(2):122.

Yarbus, A., Haigh, B., and Rigss, L. (1967). *Eye movements and vision*, volume 2. Plenum press New York.

Zhang, L., Nejati, H., Foo, L., Ma, K. T., Guo, D., and Sim, T. (2013). A talking profile to distinguish identical twins. In *Proceedings of the 10th international conference on Automatic Face and Gesture Recognition*. IEEE.

Zhao, Q. and Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3).

Appendices

Appendix A

VIP Questionnaires

This questionnaire will be presented in a computer program for ease of collection, recording and analysis.

A.1 Demographic profile

Demographic Profile (Version 1. 16 Oct 2012)

Please provide the following information for the experiment:

Gender:

Age:

Ethnicity/Race:

Religion:

Highest education qualification:

Field of studies/work:

Monthly income:

Monthly expenditures:

Monthly non-essential expenditures, e.g. movies, luxury goods:

Place of Birth:

Nationality:

A.2 Personality types

Simplified Personality Test (Version 1. 17th Oct 2012)
Answer these questions as you would usually feel or act. There are no ‘‘right’’ or ‘‘wrong’’ answers.

1. Do you prefer to draw energy from action: you tend to act, then reflect, then act further. If you are inactive, your motivation tends to decline. To rebuild your energy, you need breaks from time spent in reflection; (E)

OR

Do you prefer to expend energy through action: you prefer to reflect, then act, then reflect again. To rebuild your energy, you need quiet time alone, away from activity (I)?

[E] or [I]?

2. Are you more likely to trust information that is in the present, tangible, and concrete: that is, information that can be understood by the five senses? You tend to distrust hunches, which seem to come "out of nowhere". You prefer to look for details and facts. For you, the meaning is in the data. (S)

OR

Do you tend to trust information that is more abstract or theoretical, that can be associated with other information (either remembered or discovered by seeking a wider context or pattern). You may be more interested in future possibilities. For you, the meaning is in the underlying theory and principles which are manifested in the data. (N)

[S] or [N]?

3. Do you tend to decide things from a more detached standpoint, measuring the decision by what seems reasonable, logical, causal, consistent, and matching a given set of rules. You usually have trouble interacting with people who are inconsistent or illogical, and tend to give very direct feedback to others. You are concerned with the truth and view it as more important than being tactful. (T)

OR

Do you tend to come to decisions by associating or empathizing with the situation, looking at it 'from the inside' and weighing the situation to achieve, on balance, the greatest harmony, consensus and fit, considering the needs of the people involved. (F)

```
[T] or [F]?
```

Appendix B

Publications

B.1 Published

- Ma, K.-T., Sim, T., and Kankanhalli, M. (2013). VIP: A unifying framework for computational eye-gaze research. In `4th International Workshop on Human Behavior Understanding`. In conjunction with ACM Multimedia 2013. Springer.
- Zhang, L., Nejati, H., Foo, L., Ma, K. T., Guo, D., and Sim, T. (2013). A talking profile to distinguish identical twins. In `Proceedings of the 10th International Conference on Automatic Face and Gesture Recognition`. IEEE. `Best paper honorable mention`
- Zhang, L., Ma, K. T., Nejati, H., Foo, L., Ma, K. T., Guo, D., and Sim, T. (2013). A talking profile to distinguish identical twins. `Journal of Image and Vision Computing`.

B.2 Under review

- Ma, K.-T., Xu, Q., Sim, T., Kankanhalli, M., Li, L., and Lim, R. (2015). Eye-2-I: Using Eye-gaze for Instant Implicit profiling from eye-gaze. In `ACM CHI Conference on Human Factors in`

Computing Systems. ACM.

- Ma, K-T., Sim, T., and Kankanhalli, M. (2014). Profiling of Viewers' Demographic, Personality, Emotions and Interests from Eye-Tracking Data when Viewing Images and/or Videos. Invention Disclosure. ILO Ref: 14269A. National University of Singapore.

B.3 In preparation

- Ma, K.-T., Sim, T., and Kankanhalli, M.. VIP: A unifying framework for computational eye-gaze research. ACM Transaction on Multimedia Computing, Communications and Applications.

B.4 Website

http://mmas.comp.nus.edu.sg/VIP.html